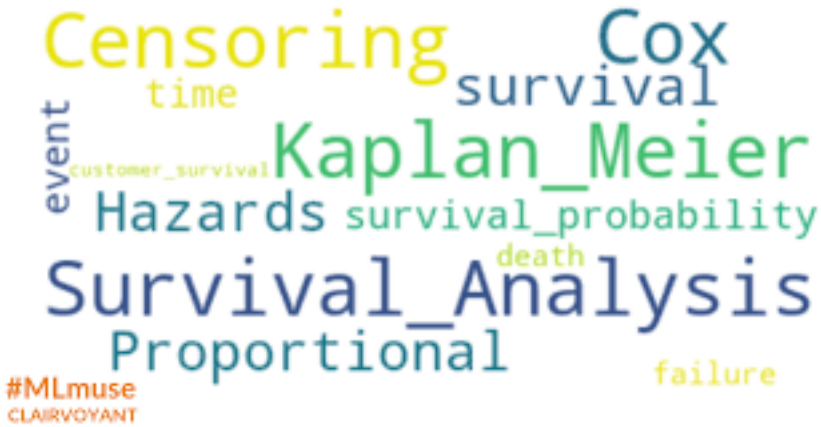
SURVIVAL ANALYSIS IN R

BY : Kikonyogo Steven



WHAT IS SURVIVAL ANALYSIS

Survival analysis focuses on describing for a given individual or group of individuals, a defined point of event called ***the failure*** (such as occurrence of a disease, cure from a disease, death, relapse after response to treatment. . .) that occurs after a period of time called ***failure time*** during which individuals are observed. It is usually expressed through the ***survival probability*** which is the probability that the event of interest has not occurred by a duration, t.

In short, survival data can be described as having the following three characteristics:

1. the dependent variable or response is the waiting time until the occurrence of a well-defined event,
2. observations can be censored, in the sense that for some units the event of interest has not occurred at the time the data are analyzed, and
3. there are predictors or explanatory variables whose effect on the waiting time we wish to assess or control.

About the project

The data set is of adjuvant chemotherapy for 888 colon cancer patients. We will focus on the study effect of treatment types on time to death. We shall visualize the data with kaplan-meier estimate of the survival function and the corresponding 95% confidence intervals. We shall fit a cox hazards proportional model, compare survival experiences and test their significance using the log rank test.You can find here the link to the data set.

Loading packages

```
if(!require(pacman)) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(
              gt,
              rio,
              here,
              dplyr,
              ggplot2,
              magrittr,
              janitor,
              survival,
              plotrix,
              flextable
              )
```

Importing data into R

```
DF <- rio::import(here("C:/Users/steven/Desktop/PROJECTS/colon.csv"))
  DF %>%
  select(
         time,
         age) %>%
  mutate(
         follow_up=time/365) %>%
  summary()
```

```
##       time             age          follow_up
##  Min.   :   8.0   Min.   :18.00   Min.   :0.02192
##  1st Qu.: 379.8   1st Qu.:53.00   1st Qu.:1.04041
##  Median :1556.0   Median :61.00   Median :4.26301
##  Mean   :1410.4   Mean   :59.81   Mean   :3.86399
##  3rd Qu.:2289.8   3rd Qu.:69.00   3rd Qu.:6.27329
##  Max.   :3329.0   Max.   :85.00   Max.   :9.12055
```

```
 DF1 <- subset(DF,rx >= 2)
```

Data Exploration

```
colon <- DF1 %>%
     select(
         time,
         status,
         rx,
```

```
        age,
        sex) %>%
    mutate(
          follow_up=time/365,
          event = ifelse(status==1, "death", "censored"),
          trt_ype  = ifelse(rx==2, "amisole", "amisole+5-FU"),
          gender = ifelse(sex == 1, "male", "female"),
          age_cat = case_when(
          age < 35 ~ "0 - 34",
          age >= 35 & age < 65 ~ "35 - 64",
          age >= 65 ~ "65+")
          )

any(is.na(colon))
```

```
## [1] FALSE
```

```
str(colon)
```

```
## 'data.frame':    594 obs. of  10 variables:
##  $ time     : int  8 9 20 23 36 40 43 45 45 49 ...
##  $ status   : int  1 1 1 0 1 1 1 1 0 1 ...
##  $ rx       : int  3 3 2 3 2 3 2 2 3 3 ...
##  $ age      : int  32 55 66 52 60 58 79 41 73 61 ...
##  $ sex      : int  0 1 1 0 0 0 1 1 1 0 ...
##  $ follow_up: num  0.0219 0.0247 0.0548 0.063 0.0986 ...
##  $ event    : chr  "death" "death" "death" "censored" ...
##  $ trt_ype  : chr  "amisole+5-FU" "amisole+5-FU" "amisole" "amisole+5-FU" ...
##  $ gender   : chr  "female" "male" "male" "female" ...
##  $ age_cat  : chr  "0 - 34" "35 - 64" "65+" "35 - 64" ...
```

```
head(colon,10)
```

```
##    time status rx age sex  follow_up    event      trt_ype gender age_cat
## 1     8      1  3  32   0 0.02191781    death amisole+5-FU female  0 - 34
## 2     9      1  3  55   1 0.02465753    death amisole+5-FU   male 35 - 64
## 4    20      1  2  66   1 0.05479452    death      amisole   male     65+
## 5    23      0  3  52   0 0.06301370 censored amisole+5-FU female 35 - 64
## 9    36      1  2  60   0 0.09863014    death      amisole female 35 - 64
## 11   40      1  3  58   0 0.10958904    death amisole+5-FU female 35 - 64
## 12   43      1  2  79   1 0.11780822    death      amisole   male     65+
## 13   45      1  2  41   1 0.12328767    death      amisole   male 35 - 64
## 14   45      0  3  73   1 0.12328767 censored amisole+5-FU   male     65+
## 15   49      1  3  61   0 0.13424658    death amisole+5-FU female 35 - 64
```

```
tabyl(colon$gender)
```

```
##  colon$gender   n   percent
##        female 298 0.5016835
##          male 296 0.4983165
```

```
tabyl(colon$trt_ype)
```

```
##  colon$trt_ype   n  percent
##        amisole 305 0.513468
##   amisole+5-FU 289 0.486532
```

```
tabyl(colon$age_cat)
```

```
##  colon$age_cat   n    percent
##         0 - 34  21 0.03535354
##        35 - 64 339 0.57070707
##            65+ 234 0.39393939
```

```
tabyl(colon$event)
```

```
##  colon$event   n  percent
##     censored 309 0.520202
##        death 285 0.479798
```

```
colon %>%
  tabyl(gender, age_cat) %>%
  adorn_totals(where = "both") %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>%
  gt()
```

gender	0 - 34	35 - 64	65+	Total
female	12 (4.0%)	168 (56.4%)	118 (39.6%)	298 (100.0%)
male	9 (3.0%)	171 (57.8%)	116 (39.2%)	296 (100.0%)
Total	21 (3.5%)	339 (57.1%)	234 (39.4%)	594 (100.0%)

```
colon %>%
  tabyl(gender, event) %>%
  adorn_totals(where = "both") %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>%
  flextable()
```

gender	censored	death	Total
female	149 (50.0%)	149 (50.0%)	298 (100.0%)
male	160 (54.1%)	136 (45.9%)	296 (100.0%)

gender	censored	death	Total
Total	309 (52.0%)	285 (48.0%)	594 (100.0%)

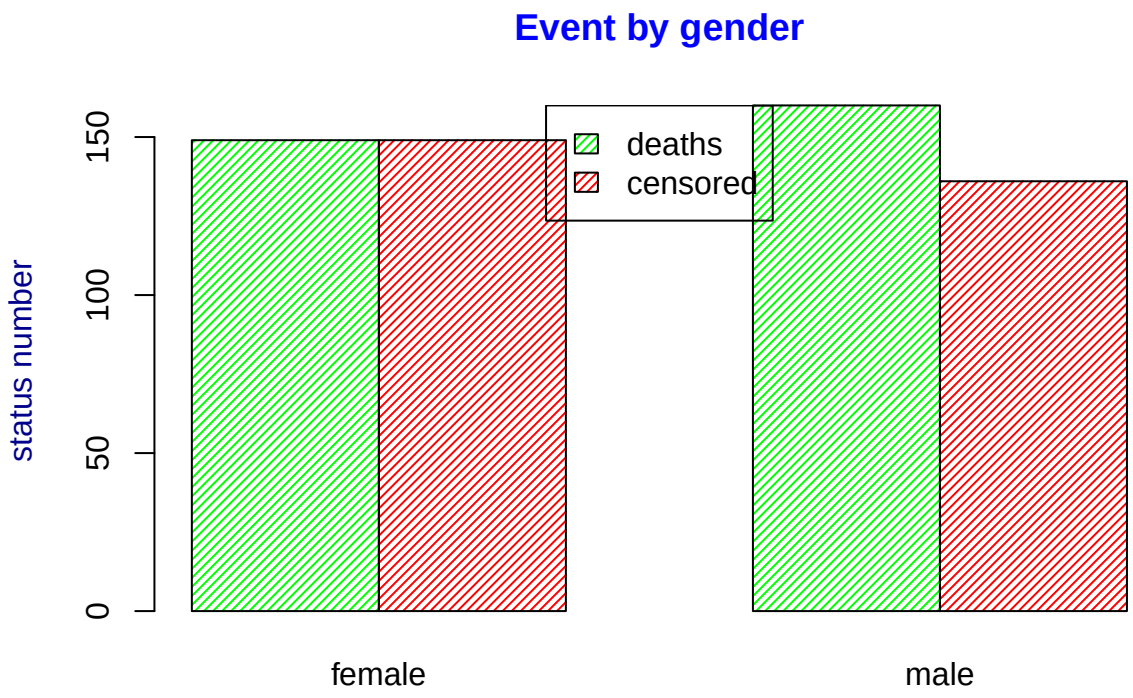
```
colon %>%
  tabyl(age_cat, event) %>%
  adorn_totals(where = "both") %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>%
  gt()
```

age_cat	censored	death	Total
0 - 34	9 (42.9%)	12 (57.1%)	21 (100.0%)
35 - 64	174 (51.3%)	165 (48.7%)	339 (100.0%)
65+	126 (53.8%)	108 (46.2%)	234 (100.0%)
Total	309 (52.0%)	285 (48.0%)	594 (100.0%)

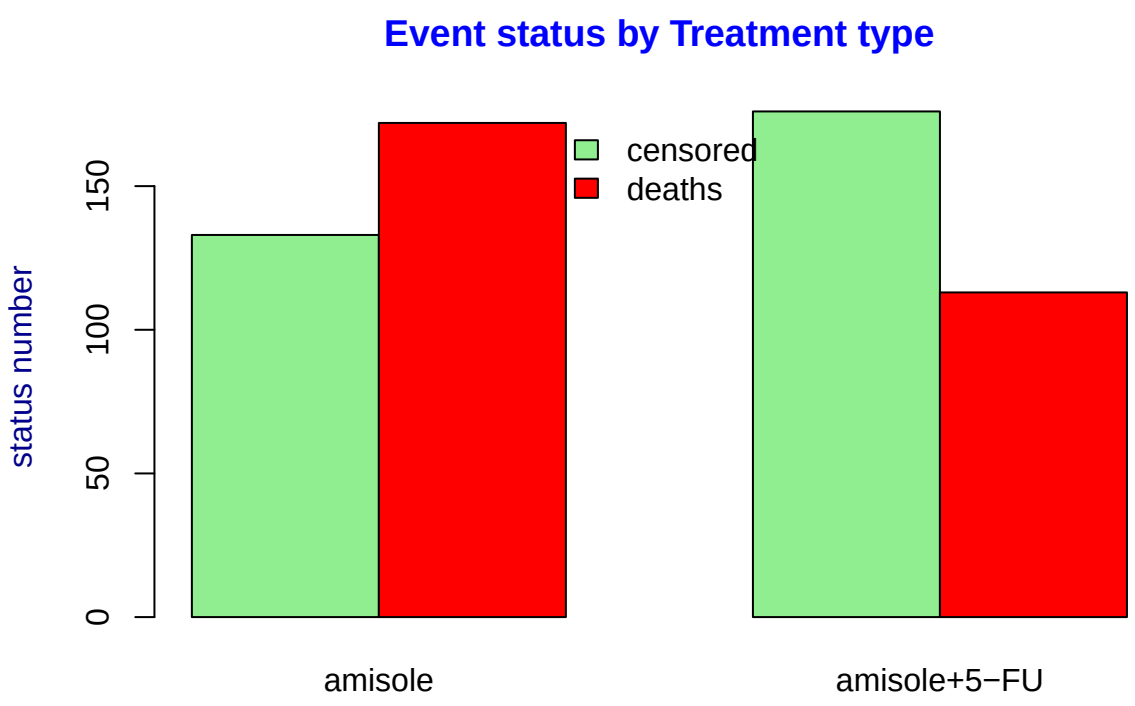
```
colon %>%
  tabyl(trt_ype, event) %>%
  adorn_totals(where = "both") %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>%
  flextable()
```

trt_ype	censored	death	Total
amisole	133 (43.6%)	172 (56.4%)	305 (100.0%)
amisole+5-FU	176 (60.9%)	113 (39.1%)	289 (100.0%)
Total	309 (52.0%)	285 (48.0%)	594 (100.0%)

```
 table1 <- table(colon$event, colon$gender)
barplot(table1, beside = T,
        main = "Event by gender",
        ylab = "status number",
        col.main = "blue",
        col.lab = "darkblue",
        col = c("green","red"),
        density = 30,
        angle = 45)
legend("top",
       legend =c("deaths","censored"),
       density = 30,angle = 45,
     fill = c("green","red"))
```



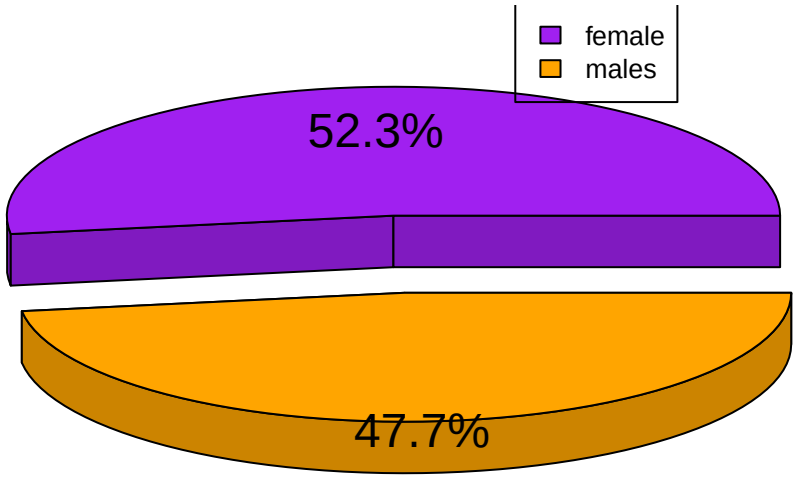

```
table2 <- table(colon$event, colon$trt_ype)
barplot(table2,beside = T,
        main ="Event status by Treatment type",
        ylab = "status number",
        col.main = "blue",
        col.lab = "darkblue",
        col = c("lightgreen","red") )
legend("top",
       legend =c("censored","deaths"),
       bty ="n",
       fill = c("lightgreen","red"))
```




```
colon3 <- colon %>%
  select(gender, event) %>%
  filter(event == "death")

table3 <- table(colon3)
piepercent <-paste0(round(100 * table3/sum(table3), 1), "%")

pie3D(table3,radius = 1.5,
                explode = 0.3,
                labels = piepercent,
                col = c("purple","orange"),
                main = "Deaths by gender ",
                col.main="blue")
legend("topright",
       c("female","males"),
       cex = 0.8,
       fill =c("purple","orange") )
```



Descriptive statistics summary

Of the 888 patients, 594 were randomly allocated to the two treatments(New treatment and old treatment) and of these, 298 were females(50.2%) and 296 males(49.8%).

The minimum age of patients was 18 years, maximum was 85 years and mean was 59.8years.The mean follow-up was 4.2years while the median follow-up was 3.9 years the maximum follow-up being 9.1years.

305(51.3%) received old treatment while 289(48.7%) received new treatment.

At the end of the study, there were 285 deaths(48%) and 309 censored(52%).Of the deaths, 149(52.3%) were females and 136(47.7%) males.Majority of the deaths were from the old treatment(56.4%) compared to new treatment(39.1%)

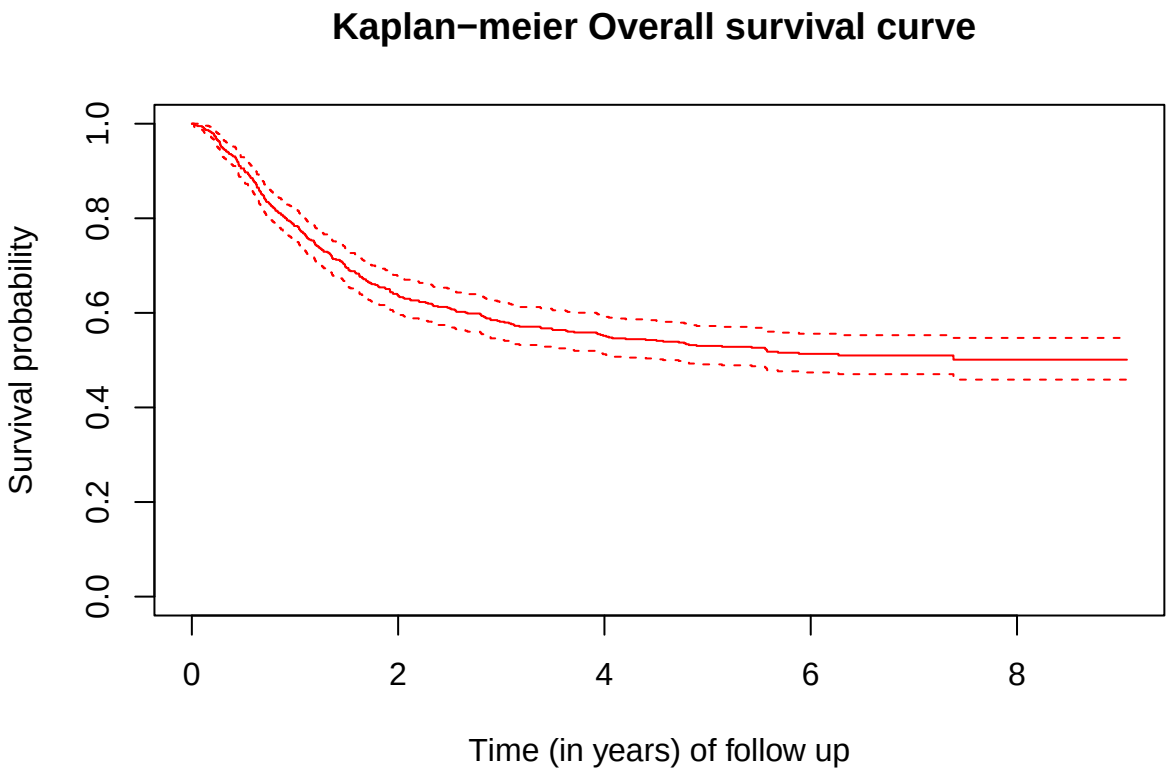
Fitting Overall survival curve

```
survobj <- Surv(time = colon$follow_up,
                event = colon$status)

 head(survobj, 10)
```

```
##  [1] 0.02191781  0.02465753  0.05479452  0.06301370+ 0.09863014  0.10958904
##  [7] 0.11780822  0.12328767  0.12328767+ 0.13424658
```

```
fitkm <- survfit(survobj ~ 1 )
plot(fitkm,
     xlab = "Time (in years) of follow up",
     ylab = "Survival probability",
     main = "Kaplan-meier Overall survival curve ",
     col="red"
     )
```



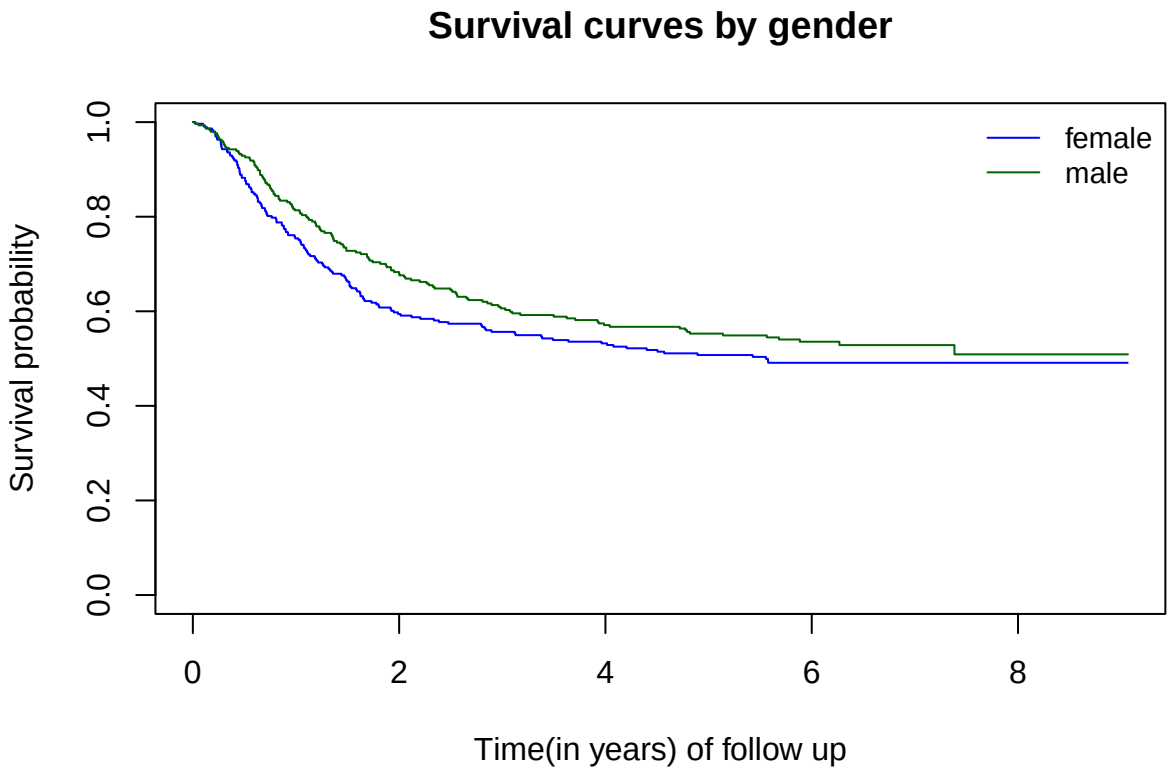
Compare survival between groups

```
colo <- c("blue", "darkgreen")

colon_gender <- survfit(Surv
                        (follow_up,status) ~sex,
                        data =colon)

 #Survival curves for male and female
 plot(
   colon_gender,
   col = colo,
   xlab = "Time(in years) of follow up",
   ylab = "Survival probability",
   main = "Survival curves by gender"
```

```
  )
 legend(
   "topright",
 legend = c("female", "male"),
 col = colo,
  lty = 1,
  cex = .9,
  bty = "n"
  )
```




```
 #Test for difference in survival by gender
coxfit_gender <- coxph(Surv
                       (follow_up,status)~gender,data=colon)
summary(coxfit_gender)
```

```
## Call:
## coxph(formula = Surv(follow_up, status) ~ gender, data = colon)
##
##   n= 594, number of events= 285
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## gendermale -0.1576    0.8542   0.1186 -1.329    0.184
##
##            exp(coef) exp(-coef) lower .95 upper .95
## gendermale    0.8542      1.171     0.677     1.078
```

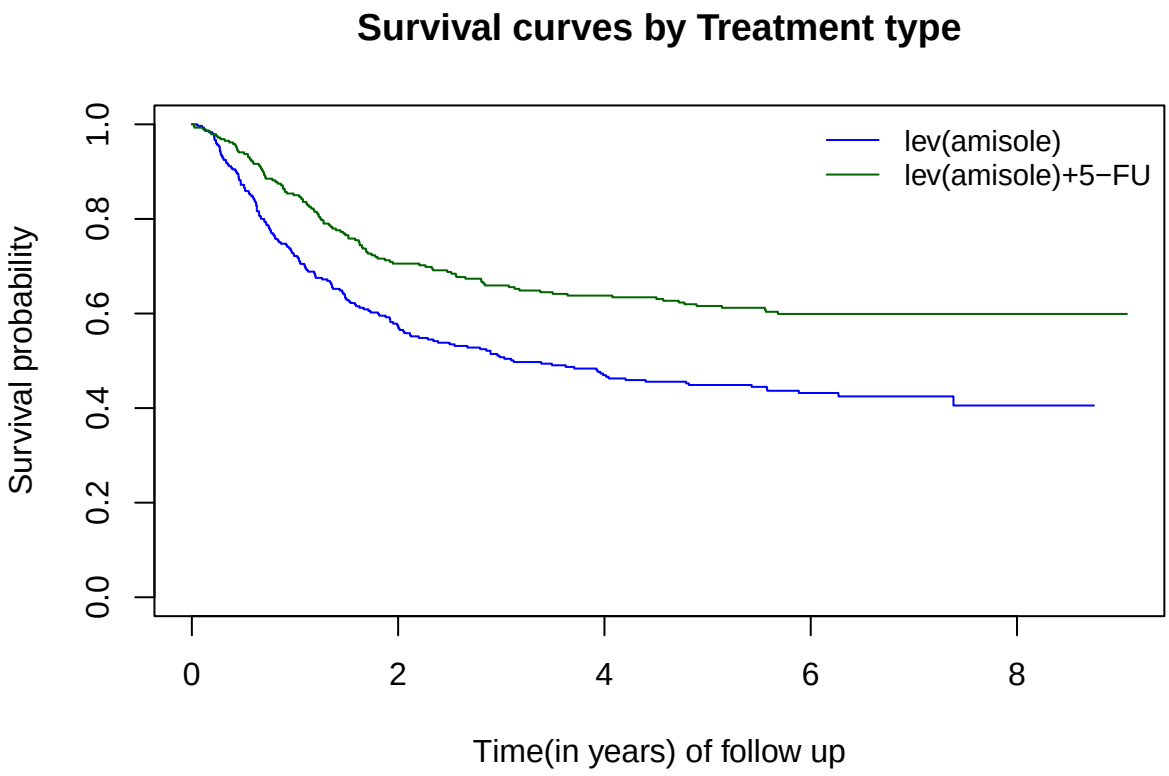
```
##
## Concordance= 0.526  (se = 0.015 )
## Likelihood ratio test= 1.77  on 1 df,   p=0.2
## Wald test            = 1.77  on 1 df,   p=0.2
## Score (logrank) test = 1.77  on 1 df,   p=0.2
```

```
survdiff(Surv
        (follow_up,status) ~ gender,data = colon)
```

```
## Call:
## survdiff(formula = Surv(follow_up, status) ~ gender, data = colon)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=female 298      149      138     0.912      1.77
## gender=male   296      136      147     0.854      1.77
##
##  Chisq= 1.8  on 1 degrees of freedom, p= 0.2
```

```
# Survival curves for  two treatment types
colon_trt <- survfit(
                  Surv(follow_up,status) ~
                    trt_ype,data =colon)

plot(
  colon_trt,
  col = colo,
  xlab = "Time(in years) of follow up",
  ylab = "Survival probability",
  main = "Survival curves by Treatment type")
legend(
  "topright",
  legend = c("lev(amisole)","lev(amisole)+5-FU"),
  col = colo,
  lty = 1,
  cex = .9,
  bty = "n"
  )
```




```
#Test for difference in survival between treatments
coxfit_trt <- coxph(
              Surv(follow_up,status)~trt_ype,
              data=colon)
summary(coxfit_trt)
```

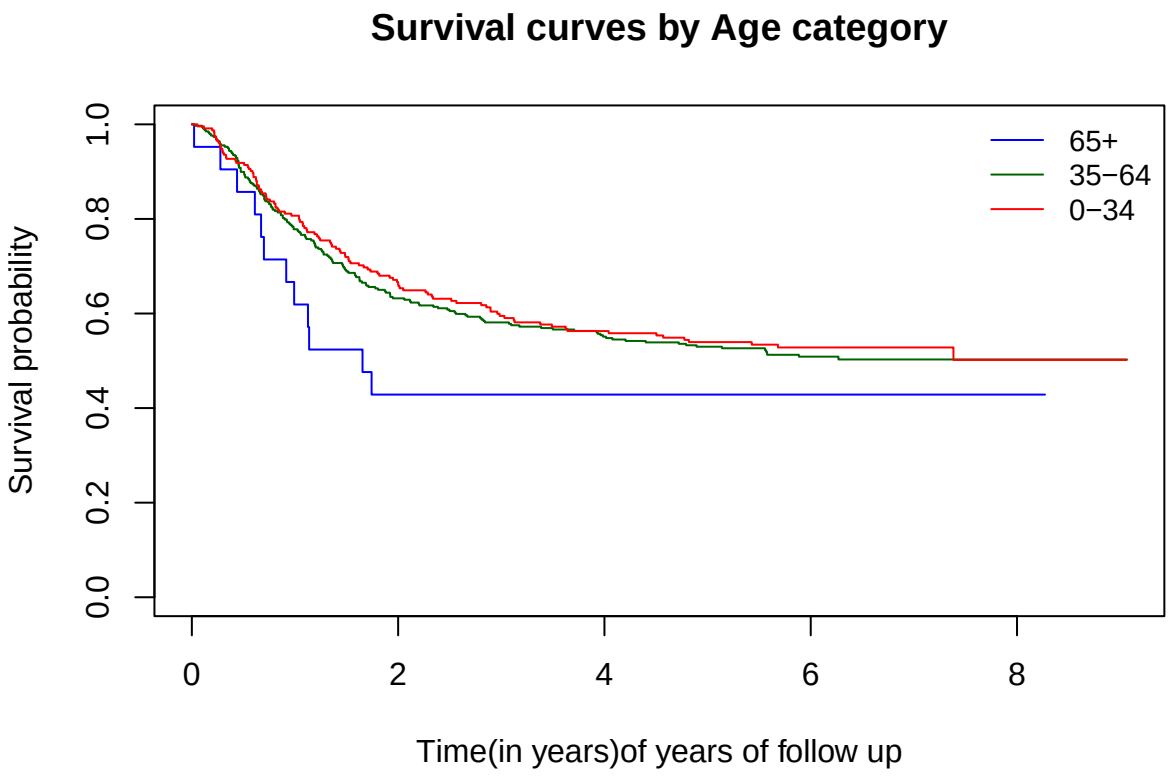
```
## Call:
## coxph(formula = Surv(follow_up, status) ~ trt_ype, data = colon)
##
##   n= 594, number of events= 285
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## trt_ypeamisole+5-FU -0.5242    0.5921   0.1212 -4.324 1.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## trt_ypeamisole+5-FU    0.5921      1.689    0.4668    0.7508
##
## Concordance= 0.566  (se = 0.015 )
## Likelihood ratio test= 19.17  on 1 df,   p=1e-05
## Wald test            = 18.69  on 1 df,   p=2e-05
## Score (logrank) test = 19.12  on 1 df,   p=1e-05
```

```
survdiff(Surv
        (follow_up,status) ~ trt_ype,data = colon)
```

```
## Call:
## survdiff(formula = Surv(follow_up, status) ~ trt_ype, data = colon)
##
##                          N Observed Expected (O-E)^2/E (O-E)^2/V
## trt_ype=amisole        305      172      135     10.02      19.1
## trt_ype=amisole+5-FU 289      113      150      9.05      19.1
##
##  Chisq= 19.1  on 1 degrees of freedom, p= 1e-05
```

```
#Survival curves for age categories
colon_age_cat <- survfit(
                 Surv(follow_up,status)~age_cat,data=colon)
col_age <- c("blue", "darkgreen", "red")

plot(
  colon_age_cat,
  col = col_age,
  xlab = "Time(in years)of years of follow up",
  ylab = "Survival probability",
  main = "Survival curves by Age category")
legend(
  "topright",
  legend = c("65+","35-64","0-34"),
  col = col_age,
  lty = 1,
  cex = .9,
  bty = "n"
  )
```




```
#Test for difference in survival in age categories
coxfit_age_cat <- coxph(
                  Surv(follow_up,status) ~ age_cat,
                  data=colon)
summary(coxfit_age_cat)
```

```
## Call:
## coxph(formula = Surv(follow_up, status) ~ age_cat, data = colon)
##
##   n= 594, number of events= 285
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## age_cat35 - 64 -0.3475    0.7064   0.2991 -1.162    0.245
## age_cat65+     -0.4048    0.6671   0.3044 -1.330    0.184
##
##                exp(coef) exp(-coef) lower .95 upper .95
## age_cat35 - 64    0.7064      1.416    0.3931     1.270
## age_cat65+        0.6671      1.499    0.3674     1.211
##
## Concordance= 0.516  (se = 0.015 )
## Likelihood ratio test= 1.63  on 2 df,   p=0.4
## Wald test            = 1.79  on 2 df,   p=0.4
## Score (logrank) test = 1.81  on 2 df,   p=0.4
```

```
survdiff(
      Surv(follow_up,status) ~ age_cat,data = colon)
```

```
## Call:
## survdiff(formula = Surv(follow_up, status) ~ age_cat, data = colon)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## age_cat=0 - 34   21       12     8.39    1.5481     1.597
## age_cat=35 - 64 339      165   163.37    0.0162     0.038
## age_cat=65+     234      108   113.23    0.2417     0.401
##
##  Chisq= 1.8  on 2 degrees of freedom, p= 0.4
```

- ***Gender :***There is no statistically significant difference in survival between males and females(p=0.2). The HR=0.85, indicates that patients who are male have a reduced risk of death(by 15%) compared to females.The 95% C.I is (0.677-1.08) indicating that HR is not statistically different from 1*.
- ***AGE CATEGORY :*** *There is no statistically significant difference in survival among the age categories(p=0.4).The HR for 35-64(HR=0.71) indicates that patients in this category have a reduced risk of death(by 29%) compared to age category 0-34.The 95% C.I is (0.393-1.27) indicating that HR is not statistically different from 1. The HR for 65+(HR=0.67) indicates that patients in this category have a reduced risk of death(by 33%) compared to age category 0-34.The 95% C.I is (0.37-1.21) indicating that HR is not statistically different from 1*
- ***TREATMENT TYPES :*** *There is a statistically significant difference in survival between the old and new treatment(p<0.05). The HR=0.6, indicates that patients who received the new treatment have a reduced risk of death(by 40%) compared to those who received old treatment.The 95% C.I is (0.47-0.75) indicating that HR is statistically different from 1.*

Cox Regression analysis

```
colon_cox_sexagecat <- coxph(
                         Surv(follow_up,status) ~
                           gender + age_cat,
                         data = colon
                         )

colon_cox_sexagecat
```

```
## Call:
## coxph(formula = Surv(follow_up, status) ~ gender + age_cat, data = colon)
##
##                     coef exp(coef) se(coef)      z     p
## gendermale      -0.1573    0.8544   0.1187 -1.325 0.185
## age_cat35 - 64 -0.3362    0.7144   0.2992 -1.124 0.261
## age_cat65+     -0.4000    0.6703   0.3044 -1.314 0.189
##
## Likelihood ratio test=3.39  on 3 df, p=0.3359
## n= 594, number of events= 285
```

```
test.colon_cox_sexagecat <-cox.zph(colon_cox_sexagecat)
test.colon_cox_sexagecat
```

```
##         chisq df     p
## gender   3.63  1 0.057
## age_cat  1.82  2 0.404
## GLOBAL   5.40  3 0.145
```

```
colon_cox_sextrt <- coxph(
                      Surv(follow_up,status) ~
                        gender + trt_ype,
                      data = colon
                      )

colon_cox_sextrt
```

```
## Call:
## coxph(formula = Surv(follow_up, status) ~ gender + trt_ype, data = colon)
##
##                          coef exp(coef) se(coef)      z        p
## gendermale           -0.1733    0.8409   0.1187 -1.460    0.144
## trt_ypeamisole+5-FU -0.5293    0.5890   0.1213 -4.365 1.27e-05
##
## Likelihood ratio test=21.3  on 2 df, p=2.365e-05
## n= 594, number of events= 285
```

```
test_colon_coxsextrt <- cox.zph(colon_cox_sextrt)
test_colon_coxsextrt
```

```
##         chisq df     p
## gender  4.178  1 0.041
## trt_ype 0.504  1 0.478
## GLOBAL  4.759  2 0.093
```

```
#fit the model
colon_surv_cox <-  coxph(
                   Surv(follow_up, status) ~
                     gender + age + trt_ype,
                   data = colon
                   )
 summary(colon_surv_cox)
```

```
## Call:
## coxph(formula = Surv(follow_up, status) ~ gender + age + trt_ype,
##     data = colon)
##
##   n= 594, number of events= 285
##
##                           coef exp(coef)  se(coef)      z Pr(>|z|)
## gendermale           -0.174809  0.839617  0.118648 -1.473    0.141
## age                  -0.007515  0.992513  0.004919 -1.528    0.127
```

```
## trt_ypeamisole+5-FU -0.522006  0.593329  0.121339 -4.302 1.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## gendermale             0.8396      1.191    0.6654    1.0594
## age                    0.9925      1.008    0.9830    1.0021
## trt_ypeamisole+5-FU    0.5933      1.685    0.4677    0.7526
##
## Concordance= 0.588  (se = 0.017 )
## Likelihood ratio test= 23.6  on 3 df,   p=3e-05
## Wald test            = 23.24  on 3 df,   p=4e-05
## Score (logrank) test = 23.68  on 3 df,   p=3e-05
```

```
#test the proportional hazard model
colon_surv_cox_ph_test <- cox.zph(colon_surv_cox)
colon_surv_cox_ph_test
```

```
##         chisq df     p
## gender  4.129  1 0.042
## age     0.831  1 0.362
## trt_ype 0.534  1 0.465
## GLOBAL  5.496  3 0.139
```