SURVIVAL ANALYSIS IN R

BY : Kikonyogo Steven

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 594 colon cancer patients. We will focus more on the study effect of new treatment type in comaparison with the old treatment type in regards to time to death from colon cancer. 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.

(i) Importing data and loading packages

```
if(!require(pacman)) install.packages("pacman")
pacman::p_load(broom,kableExtra,tidyverse,survival,survminer, gtsummary)

colon <- read.csv("colon.csv")
```

(ii) Data manipulation and Descriptive statistics

```
colon1 <- colon %>%
  select(
    time,
    status,
   Treatment.type,
    sex,
    age,
    adhere,
    postive.nodes,
   age.group) %>%
```

```
  mutate(
    time=time/31)

colon1$adhere <- factor(colon1$status, levels = c("0","1"),
                        labels = c("No","Yes"))
colon1$sex <- factor(colon1$sex, levels = c("0","1"),
                        labels = c("male","female"))

colon1$Treatment.type <- factor(colon1$Treatment.type, levels = c("2","3"),
                        labels = c("old treatment","New treatment"))

colon1$status <- factor(colon1$status, levels = c("0","1"),
                        labels = c("censored","Death"))

colon1$age.group  <- factor(colon1$age.group , levels = c("0","1"),
                        labels = c(",< = 60 yrs","> 60 yrs"))

dim(colon)
```

```
## [1] 594   8
```

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

```
##    time status Treatment.type sex age adhere postive.nodes age.group
## 1     8      1              3   0  18      0             14         0
## 2     9      1              3   1  22      0              5         0
## 3    20      1              2   1  25      0              4         0
## 4    23      0              3   0  27      0              3         0
## 5    36      1              2   0  27      1              4         0
## 6    40      1              3   0  30      0              5         0
## 7    43      1              2   1  30      0             19         0
## 8    45      1              2   1  30      0              7         0
## 9    45      0              3   1  30      0              2         0
## 10   49      1              3   0  31      0              9         0
```

```
summary(colon1)
```

```
##       time               status          Treatment.type     sex
##  Min.   :  0.2581   censored:309   old treatment:305    male  :298
##  1st Qu.: 13.2903   Death   :285   New treatment:289    female:296
##  Median : 57.7097
##  Mean   : 46.7331
##  3rd Qu.: 74.1210
##  Max.   :106.7419
##       age        adhere    postive.nodes         age.group
##  Min.   :18.00   No :309   Min.   : 0.00   ,< = 60 yrs:267
##  1st Qu.:53.00   Yes:285   1st Qu.: 1.00   > 60 yrs   :327
##  Median :61.00             Median : 2.00
##  Mean   :59.68             Mean   : 3.64
##  3rd Qu.:68.75             3rd Qu.: 5.00
##  Max.   :85.00             Max.   :27.00
```

Table 1: Table1: Descriptive statistics

Characteristic	**N = 594**
___ median follow up time (months)___	58 (13, 74)
___mean age,yrs (sd)___	61 (53, 69)
___median positive nodes (IQR)___	2 (1, 5)
___sex___	
male	298 (50%)
female	296 (50%)
___status___	
censored	309 (52%)
Death	285 (48%)
___Treatment type___	
old treatment	305 (51%)
New treatment	289 (49%)
___age.group___	
,< = 60 yrs	267 (45%)
> 60 yrs	327 (55%)
___adhere___	
No	309 (52%)
Yes	285 (48%)

```
descriptives <-colon1 %>%
  select(time,age,postive.nodes,sex,status,Treatment.type,age.group,adhere)
tbl_summary(descriptives,
            label = list(
              time ~ " median follow up time (months)",
                Treatment.type ~ "Treatment type",
                age ~ "mean age,yrs (sd)",
                postive.nodes~ "median positive nodes (IQR)"),
            type = list(adhere ~ "categorical"))%>%
  modify_caption("Table1: Descriptive statistics") %>%
  bold_labels()
```

```
 colon1 %>%
  select(time,age,postive.nodes,sex,status, Treatment.type,age.group,adhere) %>%
  tbl_summary(by=Treatment.type,
              label = list(
              time ~ " median follow up(months,(IQR))",
              adhere ~ " Adherence to nearby organs",
              age ~ "mean age,yrs (sd)",
              postive.nodes~ "median positive nodes (IQR)"),
          statistic = list(age ~ "{mean}({sd})"),
 type = list(adhere ~ "categorical")) %>%
  modify_caption("Table2: descriptive statistics by treatment type")
```

```
 colon1 %>%
  select(time,age,postive.nodes,sex,status,Treatment.type,age.group) %>%
  tbl_summary(by=status,
              label = list(
```

Table 2: Table2: descriptive statistics by treatment type

Characteristic	**old treatment**, N = 305	**New treatment**, N = 289
median follow up(months,(IQR))	34 (10, 71)	65 (18, 79)
mean age,yrs (sd)	58(12)	62(12)
median positive nodes (IQR)	2 (1, 5)	2 (1, 4)
sex		
male	144 (47%)	154 (53%)
female	161 (53%)	135 (47%)
status		
censored	133 (44%)	176 (61%)
Death	172 (56%)	113 (39%)
age.group		
,< = 60 yrs	160 (52%)	107 (37%)
> 60 yrs	145 (48%)	182 (63%)
Adherence to nearby organs		
No	133 (44%)	176 (61%)
Yes	172 (56%)	113 (39%)

Table 3: Table2: descriptive statistics by status of last follow up

Characteristic	**censored**, N = 309	**Death**, N = 285
median follow up(months,(IQR))	73 (68, 83)	13 (7, 24)
mean age,yrs (sd)	68(7)	50(9)
median positive nodes (IQR)	2 (1, 3)	3 (2, 6)
sex		
male	149 (48%)	149 (52%)
female	160 (52%)	136 (48%)
Treatment type		
old treatment	133 (43%)	172 (60%)
New treatment	176 (57%)	113 (40%)
age.group		
,< = 60 yrs	15 (4.9%)	252 (88%)
> 60 yrs	294 (95%)	33 (12%)

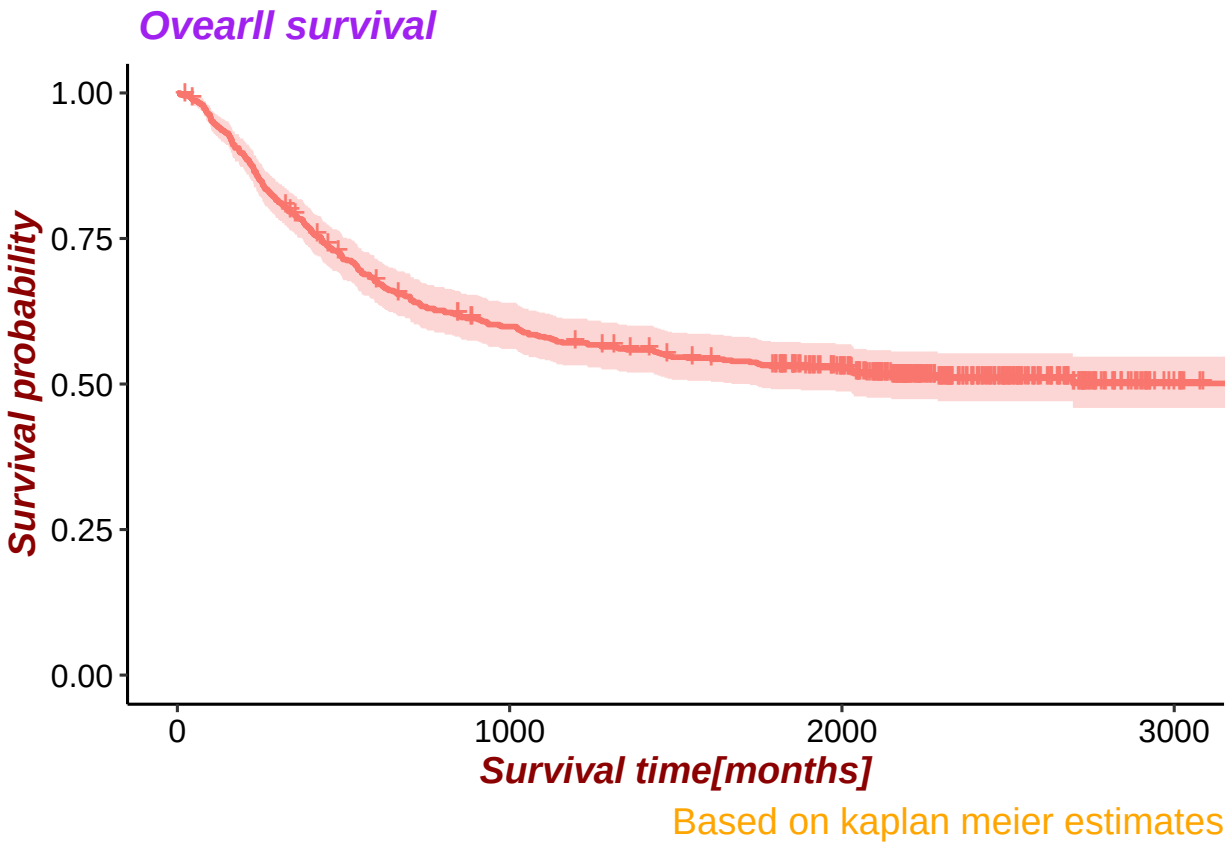
```
            time ~ " median follow up(months,(IQR))",
             Treatment.type ~ "Treatment type",
            age ~ "mean age,yrs (sd)",
            postive.nodes~ "median positive nodes (IQR)"),
        statistic = list(age ~ "{mean}({sd})")) %>%
  modify_caption("Table2: descriptive statistics by status of last follow up")
```

The mean age of patients was 59.6 years. The median follow up was 57 months. 305 patients received old treatment while 289 received new treatment. There were 285 deaths and 309 censored observations. '

(iii) Fitting a COX proportional model

```
# Fitting overall survival curve
fitkm <- survfit(Surv(time,status)~1, data = colon)
ggsurvplot(fitkm,
```

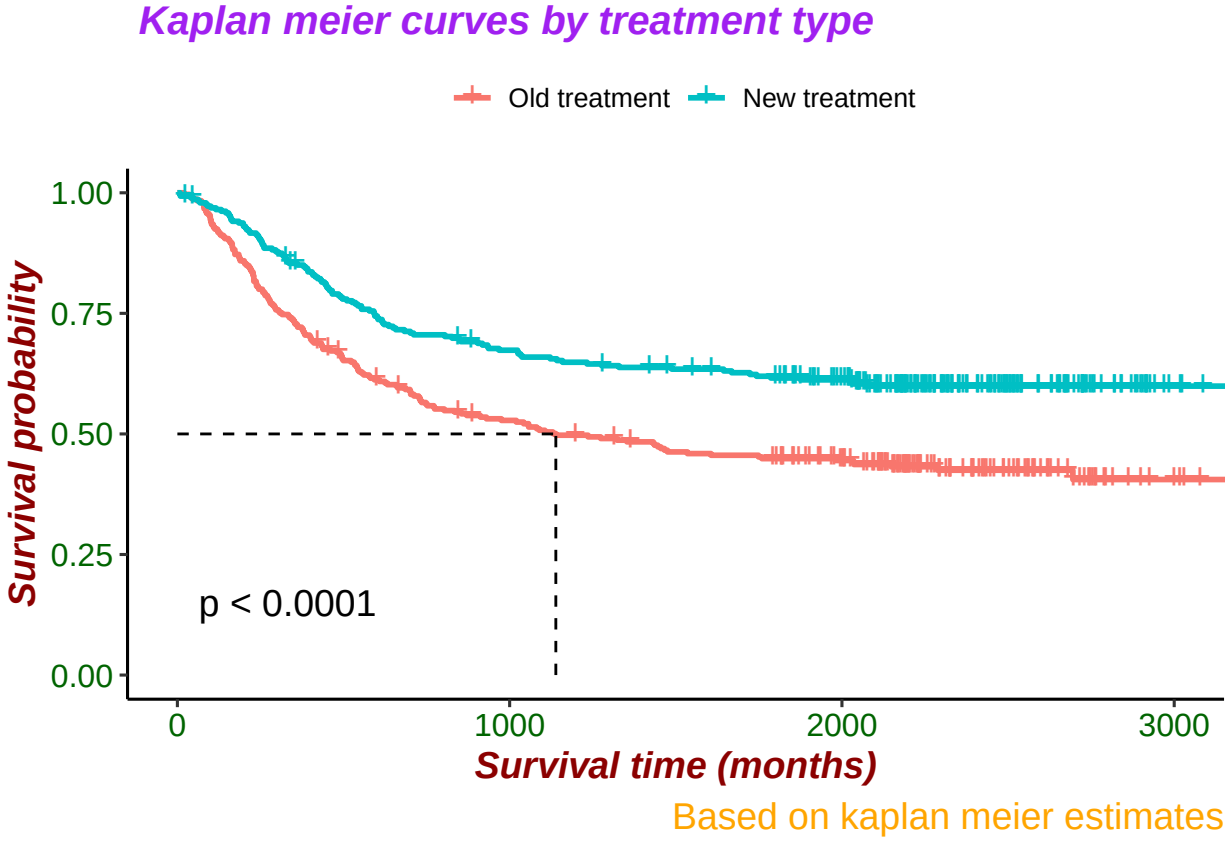
```
          xlab = "Survival time[months]",
  ylab = "Survival probability",
  submain =" Ovearll survival",
  caption = "Based on kaplan meier estimates")+
  theme_survminer(
    legend = "none",
     font.main = c(16, "bold","darkblue"),
    font.submain = c(15,"bold.italic","purple"),
    font.caption = c(14,"plain","orange"),
    font.x = c(14,"bold.italic","darkred"),
    font.y = c(14,"bold.italic","darkred"))
```




```
fitkm1 <- survfit(Surv
                        (time,status)~Treatment.type,data =colon)

#Fitting Kaplan meier  curves for Treatment types
ggsurvplot(
  fitkm1,
  surv.median.line = "hv",
  xlab = "Survival time (months)",
  ylab = "Survival probability",
  submain =" Kaplan meier curves by treatment type",
  legend.title = "",
  legend.labs = c("Old treatment","New treatment"),
  pval = TRUE,
```

```
  pval.size = 5.0,
  pval.coord = c(65,0.15),
  caption = " Based on kaplan meier estimates")+
  theme_survminer(
    font.main = c(16, "bold","darkblue"),
    font.submain = c(15,"bold.italic","purple"),
    font.caption = c(14,"plain","orange"),
    font.x = c(14,"bold.italic","darkred"),
    font.y = c(14,"bold.italic","darkred"),
    font.tickslab = c(12,"plain","darkgreen"))
```




```
# Test for statistical difference between treatment types

coxfitkm1 <- coxph(
  Surv(time,status)~Treatment.type,
  data=colon)


survdiff(Surv
         (time,status) ~ Treatment.type,data = colon)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ Treatment.type, data = colon)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
```

Table 4: **Table1:Univariate analysis for survival**

Characteristic	**N**	**Event N**	**HR**	**SE**	**95% CI**	**p-value**
postive.nodes	594	285	1.11***	0.013	1.09, 1.14	___<0.001___
sex						
0	298	149	—	—	—	
1	296	136	0.85	0.119	0.68, 1.08	0.2
Treatment.type						
2	305	172	—	—	—	
3	289	113	0.59***	0.121	0.47, 0.75	___<0.001___
adhere						
0	511	238	—	—	—	
1	83	47	1.40*	0.160	1.02, 1.92	___0.035___
age.group						
0	267	252	—	—	—	
1	327	33	0.00	1,158	0.00, Inf	>0.9

```
## Treatment.type=2 305      172      135     10.02      19.1
## Treatment.type=3 289      113      150      9.05      19.1
##
##  Chisq= 19.1  on 1 degrees of freedom, p= 1e-05
```

(iv) Univariate analysis of factors for survival

```
colon$sex  <- as.factor(colon$sex )
colon$Treatment.type  <- as.factor(colon$Treatment.type )
colon$adhere  <- as.factor(colon$adhere )
colon$age.group  <- as.factor(colon$age.group )
univariate <- tbl_uvregression(
  colon %>%
    select(time,postive.nodes,sex,status,Treatment.type,adhere, age.group),
  method = coxph,
  y=Surv(time,status),
  exponentiate = TRUE) %>%
  add_n(location="level") %>%
  add_nevent(location="level") %>%
  add_significance_stars(
    hide_ci =F ,hide_p =F ,hide_se = F) %>%
  modify_caption ("**Table1:Univariate analysis for survival**") %>%
  bold_p()
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `model = map(...)`.
## Caused by warning in `coxph.fit()`:
## ! Loglik converged before variable  1 ; coefficient may be infinite.
```

```
univariate
```

v) Multivariate analysis of factors for survival

Table 5: **Table2:Multivariable analysis for recurrence**

Predictor	**HR**	**95% CI**	**p-value**
postive.nodes	1.11***	1.09, 1.14	___<0.001___
Treatment.type			
2	—	—	
3	0.60***	0.47, 0.76	___<0.001___
adhere			
0	—	—	
1	1.33	0.97, 1.82	0.073

```
cm<- coxph(Surv(time,status )~

             postive.nodes+Treatment.type+adhere,
           data =  colon)
multivariate <- tbl_regression(cm,exponentiate = TRUE) %>%
  add_significance_stars(
    hide_ci =F ,hide_p =F ,hide_se = T) %>%
  modify_header(label="**Predictor**") %>%
  modify_caption ("**Table2:Multivariable analysis for recurrence**") %>%
  bold_p()
multivariate
```

The factors significant for survival at univariate analysis were; number of positive nodes, treatment type, and adherence to nearby organs (All $p < 0.05$). At multivariate analysis level, the factors were: Number of positive nodes, HR = 1.1 [(1.09-1.14);P<0.001] and New treatment type compared to old treatment type, HR = 0.6 [(0.47-0.76);P<0.001].