

Kaplan-Meier plots

Zhaoqi Liu

2/20/2021

```
bcdf<-readRDS("breastcancerdf.rds")
```

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
## filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
## intersect, setdiff, setequal, union
```

```
library(survival)
```

```
library(survminer)
```

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

```
## Loading required package: ggpubr
```

```
#####Pre-select Select 8 variables that are used to draw the Kaplan Meier Plots: race(DEMO_RACE),age(DEMO_AGE_AT_D  
ER(PATH_ER), PR(PATH_PR), HER2(PATH_HER2), grade(PATH_SURGERYOVERALLGRADE),  
stage, menopause(HORMO_HORMO_MENOPAUSESTATUS)  
and 5 death and relapse status variables (response) Then, we change the variable names for convenience.
```

```
kmpdf<-bcdf[,c(1:3,6:8,13,20,4,15:16,19,18,17)]
```

```
names(kmpdf)<-c("id","race","age","ER","PR","HER2","grade","stage","menopause",  
               "metastatic","survival","survival_month","relapse","relapse_month")
```

```
head(kmpdf)
```

```
## # A tibble: 6 x 14
```

```
##   id race   age   ER   PR  HER2 grade stage menopause metastatic survival  
##   <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <chr>      <dbl>  
## 1  835 White   91    1    1    0    NA    0      0 No         0      1  
## 2  838 White   87    1    0    1    0    0      0 No         0      1  
## 3  837 White   43    0    1    1    NA    0      1 No         0      0  
## 4  841 White   59    1    1    0    0    0      0 No         0      0  
## 5  865 White   69    1    1    0    0    0      0 No         0      0  
## 6  881 White   42    1    1    1    1    0      1 No         0      0
```

```
## # ... with 3 more variables: survival_month <dbl>, relapse <dbl>,  
## # relapse_month <dbl>
```

```
median(kmpdf$age)
```

```
## [1] 54
mean(kmpdf$age)

## [1] 55.58671
hist(kmpdf$age,freq=FALSE,main="histogram of age", xlab="age" )
lines(density(kmpdf$age),lwd=2, col=2) #kernel density plot
```



The density plot shows that the age variable is approximately normal distributed. Since `mean(55.58671)` is slightly greater than `median(54)`, the variable is slightly right skewed. Divide age into two groups by the median of age. If age is below the 54, we note it as “young”; otherwise “old”.

```
kmpdf$age<-ifelse(kmpdf$age<median(kmpdf$age),"young","old")
head(kmpdf)
```

```
## # A tibble: 6 x 14
##   id race age      ER    PR  HER2 grade stage menopause metastatic survival
##   <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <chr>          <dbl>
## 1  835 White old      1      1      0    NA      0      0 No      0 No      1
## 2  838 White old      1      0      1      0      0      0 No      0 No      1
## 3  837 White young    0      1      1    NA      0      0 No      1 No      0
## 4  841 White old      1      1      0      0      0      0 No      0 No      0
## 5  865 White old      1      1      0      0      0      0 No      0 No      0
## 6  881 White young    1      1      1      1      0      0 No      1 No      0
## # ... with 3 more variables: survival_month <dbl>, relapse <dbl>,
## #   relapse_month <dbl>
```

Delete observations that race are specified as “other”. We only focus on “Black” and “White” in race.

```
kmpdf<-kmpdf[kmpdf$race!="Other",]
summary(kmpdf[, -1])
```

```
##      race      age      ER      PR
## Length:341    Length:341    Min.   :0.0000    Min.   :0.0000
## Class :character    Class :character    1st Qu.:0.0000    1st Qu.:0.0000
## Mode  :character    Mode  :character    Median :1.0000    Median :1.0000
##                                     Mean  :0.7214    Mean  :0.6334
##                                     3rd Qu.:1.0000    3rd Qu.:1.0000
##                                     Max.   :1.0000    Max.   :1.0000
##
##      HER2      grade      stage      menopause
## Min.   :0.0000    Min.   :0.0000    Min.   :0.0000    Min.   :0.0000
## 1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.0000
## Median :0.0000    Median :0.0000    Median :0.0000    Median :0.0000
## Mean   :0.2053    Mean   :0.4768    Mean   :0.1877    Mean   :0.4194
## 3rd Qu.:0.0000    3rd Qu.:1.0000    3rd Qu.:0.0000    3rd Qu.:1.0000
## Max.   :1.0000    Max.   :1.0000    Max.   :1.0000    Max.   :1.0000
##      NA's      :18
##      metastatic      survival      survival_month      relapse
## Length:341      Min.   :0.0000    Min.   : 1.0    Min.   :0.0000
## Class :character    1st Qu.:0.0000    1st Qu.: 69.0    1st Qu.:0.0000
## Mode  :character    Median :0.0000    Median :118.0    Median :0.0000
##                                     Mean   :0.2845    Mean   :112.1    Mean   :0.1906
##                                     3rd Qu.:1.0000    3rd Qu.:148.0    3rd Qu.:0.0000
##                                     Max.   :1.0000    Max.   :225.0    Max.   :1.0000
##
##      relapse_month
## Min.   : 0.00
## 1st Qu.: 13.00
## Median : 37.00
## Mean   : 49.82
## 3rd Qu.: 82.00
## Max.   :166.00
##      NA's      :276
```

Convert most variables to factors except id and the response(survival, survival months, relapse, and relapse month). Not convert survival and relapse because survival analyses require them to be numeric events.

```
kmpdf[, -c(1, 11:14)] <- lapply(kmpdf[, -c(1, 11:14)], as.factor)
summary(kmpdf)
```

```
##      id      race      age      ER      PR      HER2      grade
## Min.   :    92    Black: 87    old  :177    0: 95    0:125    0:271    0   :169
## 1st Qu.:   907    White:254    young:164    1:246    1:216    1: 70    1   :154
## Median :   998                                     NA's: 18
## Mean   :   6885
## 3rd Qu.:  1092
## Max.   :2000978
##
##      stage      menopause      metastatic      survival      survival_month      relapse
## 0:277    0:198      No :290      Min.   :0.0000    Min.   : 1.0    Min.   :0.0000
## 1: 64     1:143      Yes: 51      1st Qu.:0.0000    1st Qu.: 69.0    1st Qu.:0.0000
##                                     Median :0.0000    Median :118.0    Median :0.0000
##                                     Mean   :0.2845    Mean   :112.1    Mean   :0.1906
##                                     3rd Qu.:1.0000    3rd Qu.:148.0    3rd Qu.:0.0000
##                                     Max.   :1.0000    Max.   :225.0    Max.   :1.0000
##
```

```
## relapse_month
## Min. : 0.00
## 1st Qu.: 13.00
## Median : 37.00
## Mean : 49.82
## 3rd Qu.: 82.00
## Max. : 166.00
## NA's : 276
```

```
sapply(c(11,13),function(x){table(kmpdf[,x])})
```

```
## [,1] [,2]
## 0 244 276
## 1 97 65
```

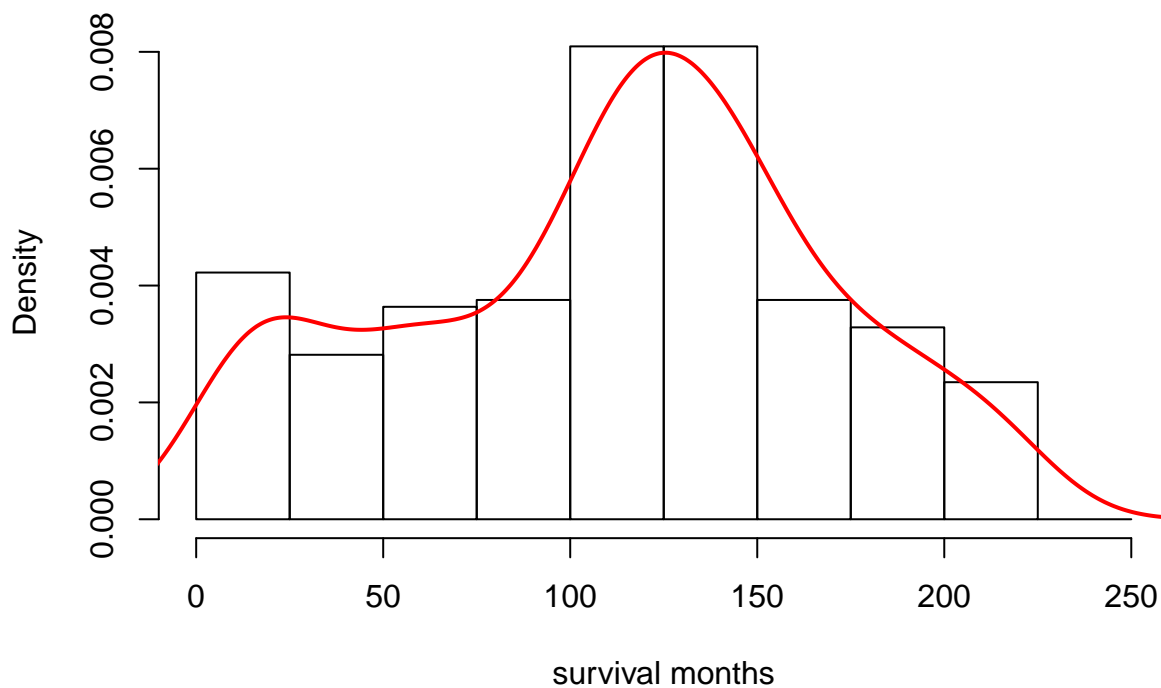
survival: 0: alive 1:dead relapse: 0: no relapse 1: local and/or distant cancer recurrence or died of disease

```
saveRDS(kmpdf,"kmpdf.rds")
```

```
###Plot the density of survival month and relapse month
```

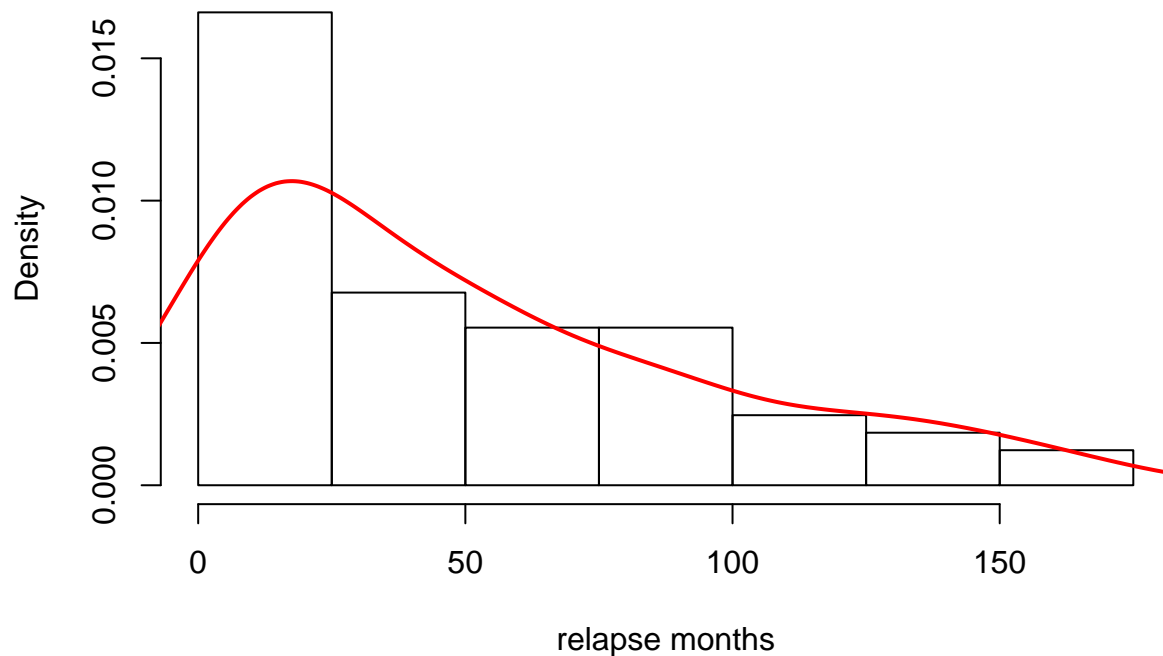
```
hist(kmpdf$survival_month,breaks =seq(0,250,25) ,freq=FALSE,xlim=c(0,250),
      main="Histogram of Survival Months",xlab= "survival months")
lines(density(kmpdf$survival_month),lwd=2,col=2)
```

Histogram of Survival Months



```
hist(kmpdf$relapse_month,breaks =seq(0,175,25) ,freq=FALSE,xlim=c(0,175),
      main="Histogram of Relapse Months",xlab= "relapse months")
lines(density(kmpdf$relapse_month,na.rm = TRUE),lwd=2,col=2)
```

Histogram of Relapse Months



The distribution of survival months is quite normal, even though it's left skewed a little bit, which means that more observations than expected have short survival months. There are only 65 observations that has the relapse months data. The histogram shows that the data maybe follow a poisson distribution with a small parameter λ .

```
attach(kmpdf)
```

Kaplan Meier Curves

####Overall Compare the survival distribution to examine whether or not there is an association between features and length of survival

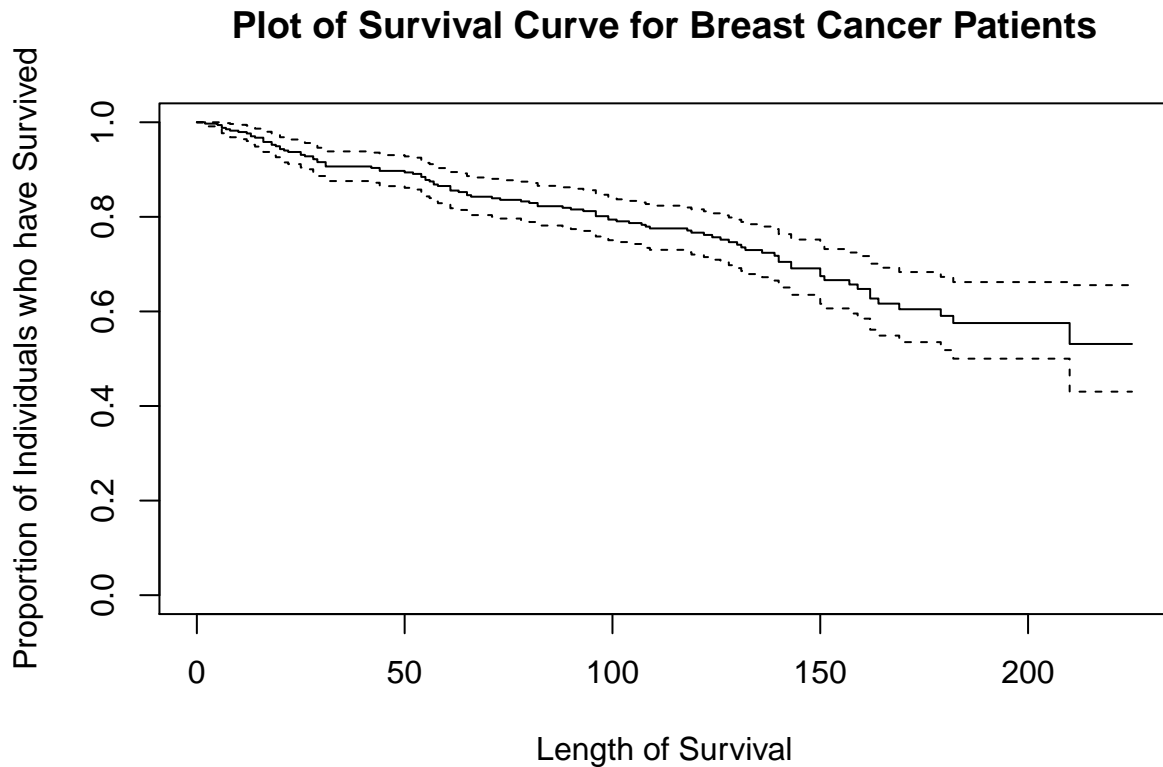
```
surv.all<-survfit(Surv(survival_month,survival)~1)
summary(surv.all)
```

```
## Call: survfit(formula = Surv(survival_month, survival) ~ 1)
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    2    340      1    0.997 0.00294    0.991    1.000
##    5    339      1    0.994 0.00415    0.986    1.000
##    6    336      2    0.988 0.00587    0.977    1.000
##    7    333      1    0.985 0.00656    0.972    0.998
##    8    332      1    0.982 0.00718    0.968    0.996
##   10    331      1    0.979 0.00774    0.964    0.995
##   12    330      1    0.976 0.00827    0.960    0.993
##   13    327      2    0.970 0.00923    0.952    0.989
##   14    324      1    0.967 0.00968    0.949    0.987
##   16    322      3    0.958 0.01090    0.937    0.980
##   18    318      2    0.952 0.01163    0.930    0.975
##   19    315      1    0.949 0.01198    0.926    0.973
```

##	20	313	2	0.943	0.01265	0.919	0.968
##	21	310	1	0.940	0.01297	0.915	0.966
##	22	308	1	0.937	0.01328	0.911	0.964
##	25	307	2	0.931	0.01388	0.904	0.959
##	26	305	1	0.928	0.01417	0.901	0.956
##	28	304	2	0.922	0.01472	0.893	0.951
##	29	300	2	0.916	0.01525	0.886	0.946
##	31	298	3	0.907	0.01599	0.876	0.938
##	42	291	1	0.903	0.01624	0.872	0.936
##	44	289	2	0.897	0.01672	0.865	0.931
##	50	282	1	0.894	0.01696	0.861	0.928
##	52	281	1	0.891	0.01719	0.858	0.925
##	54	279	2	0.884	0.01765	0.850	0.920
##	55	277	2	0.878	0.01810	0.843	0.914
##	56	275	1	0.875	0.01831	0.840	0.911
##	57	273	2	0.868	0.01873	0.832	0.906
##	58	271	1	0.865	0.01893	0.829	0.903
##	61	270	3	0.856	0.01952	0.818	0.895
##	63	265	1	0.852	0.01971	0.815	0.892
##	65	261	2	0.846	0.02009	0.807	0.886
##	66	259	1	0.843	0.02028	0.804	0.883
##	71	255	1	0.839	0.02047	0.800	0.880
##	73	252	1	0.836	0.02065	0.796	0.877
##	78	250	1	0.833	0.02084	0.793	0.874
##	80	249	1	0.829	0.02102	0.789	0.871
##	82	246	2	0.823	0.02139	0.782	0.866
##	88	241	1	0.819	0.02157	0.778	0.862
##	90	237	1	0.816	0.02175	0.774	0.859
##	93	232	1	0.812	0.02194	0.770	0.856
##	96	227	3	0.801	0.02251	0.758	0.847
##	99	223	2	0.794	0.02287	0.751	0.840
##	101	218	1	0.791	0.02306	0.747	0.837
##	104	213	1	0.787	0.02324	0.743	0.834
##	107	211	1	0.783	0.02343	0.739	0.830
##	108	208	1	0.779	0.02362	0.734	0.827
##	109	205	1	0.776	0.02381	0.730	0.824
##	118	174	1	0.771	0.02408	0.725	0.820
##	119	170	1	0.767	0.02437	0.720	0.816
##	122	161	1	0.762	0.02468	0.715	0.812
##	124	152	1	0.757	0.02502	0.709	0.807
##	126	149	1	0.752	0.02536	0.704	0.803
##	128	144	1	0.746	0.02572	0.698	0.799
##	130	136	1	0.741	0.02611	0.692	0.794
##	131	133	1	0.735	0.02650	0.685	0.789
##	132	131	1	0.730	0.02688	0.679	0.784
##	136	125	1	0.724	0.02729	0.672	0.779
##	139	115	1	0.718	0.02777	0.665	0.774
##	140	111	2	0.705	0.02874	0.651	0.763
##	143	103	2	0.691	0.02977	0.635	0.752
##	150	84	2	0.675	0.03125	0.616	0.739
##	151	80	1	0.666	0.03197	0.606	0.732
##	157	72	1	0.657	0.03284	0.596	0.725
##	159	70	1	0.648	0.03369	0.585	0.717
##	162	64	2	0.627	0.03554	0.561	0.701

##	164	58	1	0.616	0.03654	0.549	0.692
##	169	52	1	0.605	0.03771	0.535	0.683
##	179	43	1	0.591	0.03937	0.518	0.673
##	182	39	1	0.575	0.04117	0.500	0.662
##	210	13	1	0.531	0.05703	0.430	0.656

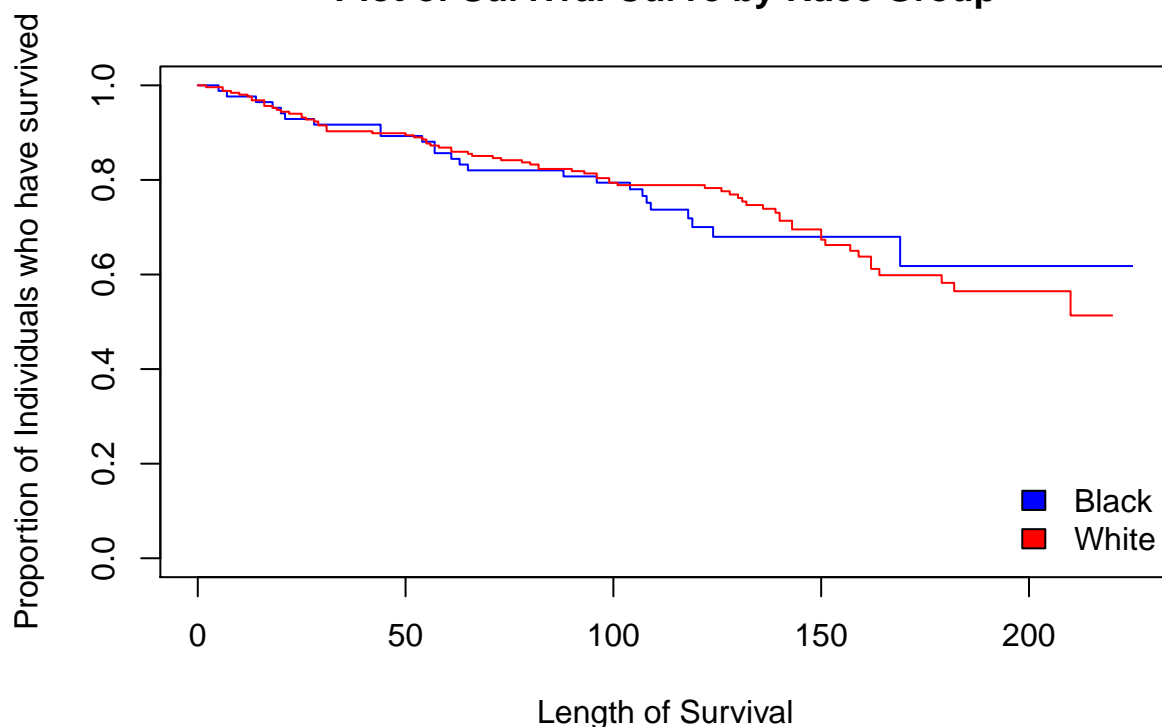
```
plot(surv.all,main="Plot of Survival Curve for Breast Cancer Patients",xlab= "Length of Survival", ylab=
```



For the 341 people in the dataset, 97 people were uncensored (followed for the entire time, until occurrence of event). Since the data has not yet dropped to 50% survival at the end of the available data, there is an NA value for median survival. The following summary goes through each time point in the study in which an individual was lost to follow up or died and re-computes the total number of people still at risk (n.risk), the number of events at that time point (n.event), the proportion of individuals who survived up until that point (survival) and the standard error (std.err) and 95% confidence interval (lower 95% CI, upper 95% CI) for the proportion of individuals who survived at that point. This plot shows the survival curve (also known as a Kaplan-Meier plot), the proportion of individual who have survived up until that particular time as a solid black line and the 95% confidence interval (the dashed lines).

```
surv.race<-survfit(Surv(survival_month,survival)~race)
plot(surv.race,col=c("blue","red"),ylim = c(0,1),
     main="Plot of Survival Curve by Race Group",
     xlab = "Length of Survival",ylab= "Proportion of Individuals who have survived")
legend("bottomright",legend=c("Black","White"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by Race Group



Race

Length of Survival

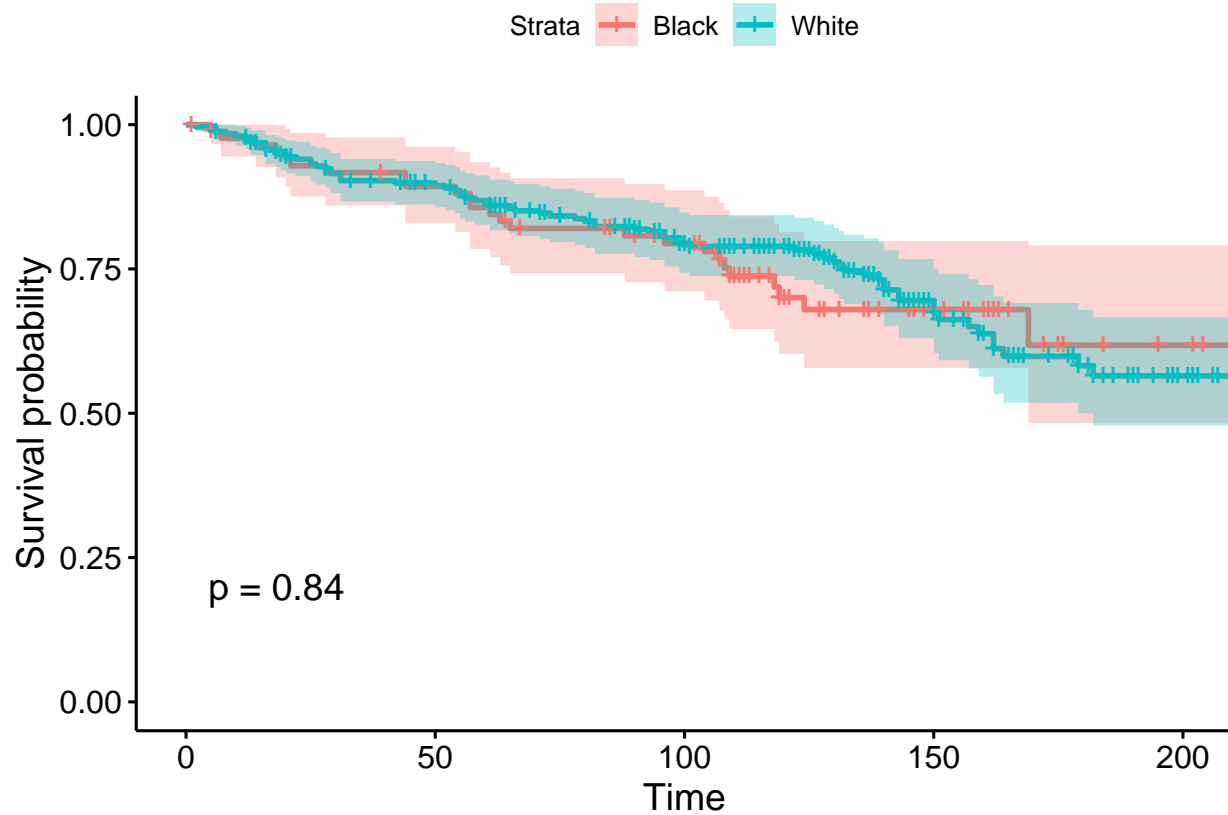
```
#Since the levels are "Black","White"
survdif(Surv(survival_month,survival)~race)
```

```
## Call:
## survdiff(formula = Surv(survival_month, survival) ~ race)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## race=Black   87      25      24.1   0.02996   0.0401
## race=White  254      72      72.9   0.00993   0.0401
##
## Chisq= 0   on 1 degrees of freedom, p= 0.8
```

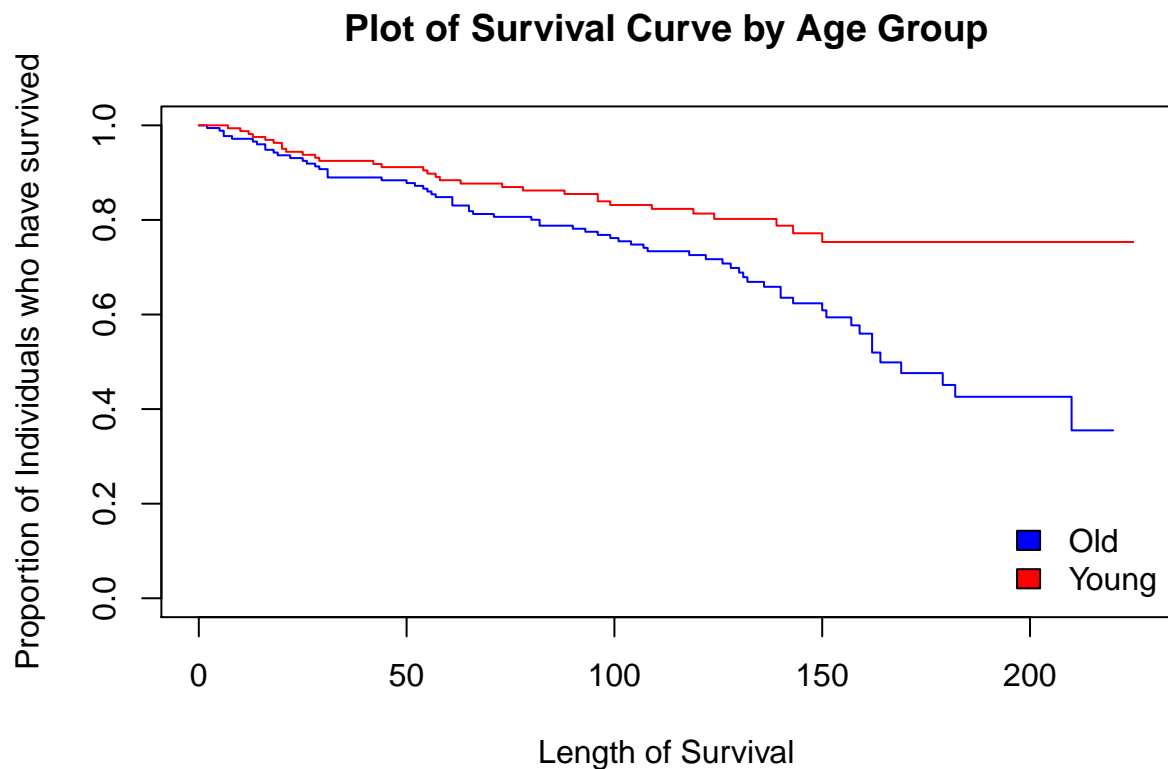
The 95% confidence interval of survival time for those on maintained chemotherapy is (, NA); NA in this case means infinity. A 95% upper confidence limit of NA/infinity is common in survival analysis due to the fact that the data is skewed.

Using `survminer` package to plot.

```
ggsurvplot(surv.race,data=kmpdf,censor.size=4,conf.int=TRUE,pval=TRUE,
            legend.labs= c("Black","White"))
```

```
surv.age<-survfit(Surv(survival_month,survival)~age)
plot(surv.age,col=c("blue","red"),ylim = c(0,1),
     main="Plot of Survival Curve by Age Group",
     xlab = "Length of Survival",ylab= "Proportion of Individuals who have survived")
legend("bottomright",legend=c("Old", "Young"),fill=c("blue", "red"),bty="n")
```



log-rank test H0: There is no difference in the survival function between those who were young and those who were old

```
survdif(Surv(survival_month,survival)~age)
```

```
## Call:
```

```
## survdif(formula = Surv(survival_month, survival) ~ age)
```

```
##
```

```
##           N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## age=old   177         66    49.7      5.35      11
```

```
## age=young 164         31    47.3      5.62      11
```

```
##
```

```
## Chisq= 11  on 1 degrees of freedom, p= 9e-04
```

```
#reject H0
```

ER estrogen receptor status

```
surv.ER<-survfit(Surv(survival_month,survival)~ER)
```

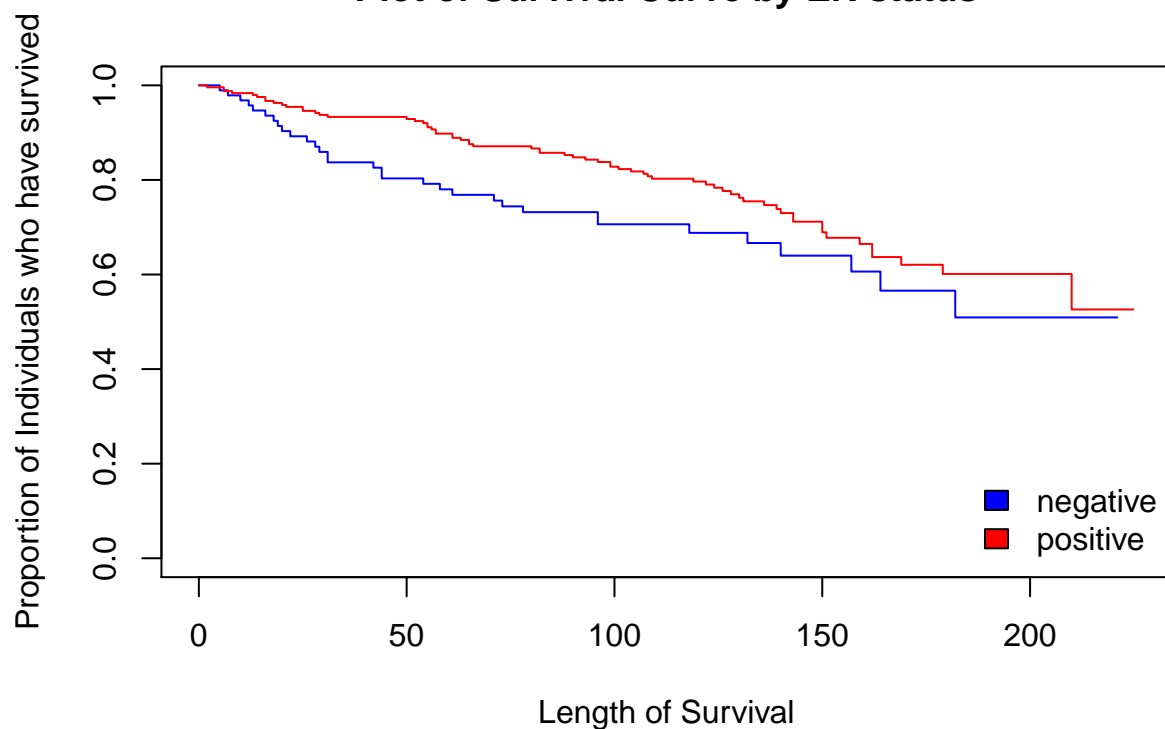
```
plot(surv.ER,col=c("blue","red"),ylim = c(0,1),
```

```
     main="Plot of Survival Curve by ER status",
```

```
     xlab = "Length of Survival",ylab= "Proportion of Individuals who have survived")
```

```
legend("bottomright",legend=c("negative","positive"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by ER status



```
#levels = 0,1 0 means negative, 1 means positive
```

```
survdif(Surv(survival_month,survival)~ER) #not reject
```

```
## Call:
```

```
## survdiff(formula = Surv(survival_month, survival) ~ ER)
```

```
##
```

```
##      N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## ER=0  95      32    24.6    2.252    3.03
```

```
## ER=1 246      65    72.4    0.764    3.03
```

```
##
```

```
## Chisq= 3 on 1 degrees of freedom, p= 0.08
```

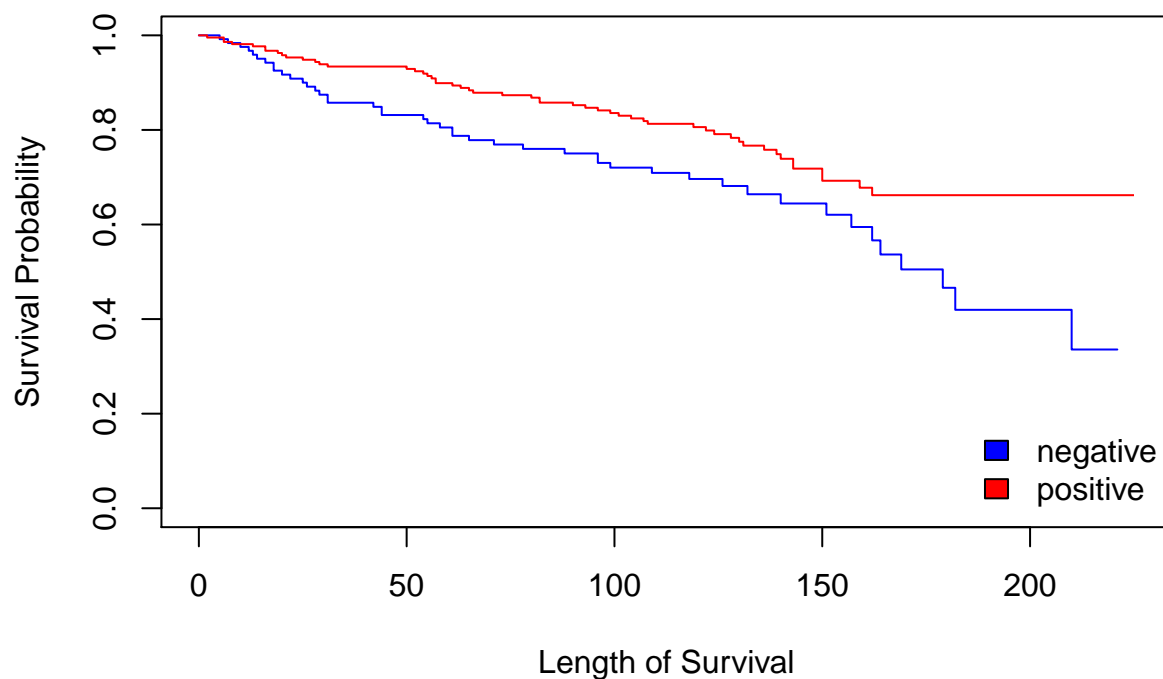
PR progesterone receptor status

```
surv.PR<-survfit(Surv(survival_month,survival)~PR)
```

```
plot(surv.PR,col=c("blue","red"),ylim = c(0,1),
     main="Plot of Survival Curve by PR status",
     xlab = "Length of Survival",ylab= "Survival Probability")
```

```
legend("bottomright",legend=c("negative","positive"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by PR status



```
#levels = 0,1 0 means negative, 1 means positive
```

```
survdif(Surv(survival_month,survival)~PR)
```

```
## Call:
```

```
## survdif(formula = Surv(survival_month, survival) ~ PR)
```

```
##
```

```
##      N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## PR=0 125      45      32.4      4.92      7.43
```

```
## PR=1 216      52      64.6      2.47      7.43
```

```
##
```

```
## Chisq= 7.4 on 1 degrees of freedom, p= 0.006
```

HER2 (human epidermal growth factor receptor 2) status positive means: When a breast cell has abnormally high levels of the HER2 gene or the HER2 protein, it is called HER2- positive. Most patients with metastatic breast cancer have HER2-negative breast cancer.

```
surv.HER2<-survfit(Surv(survival_month,survival)~HER2)
```

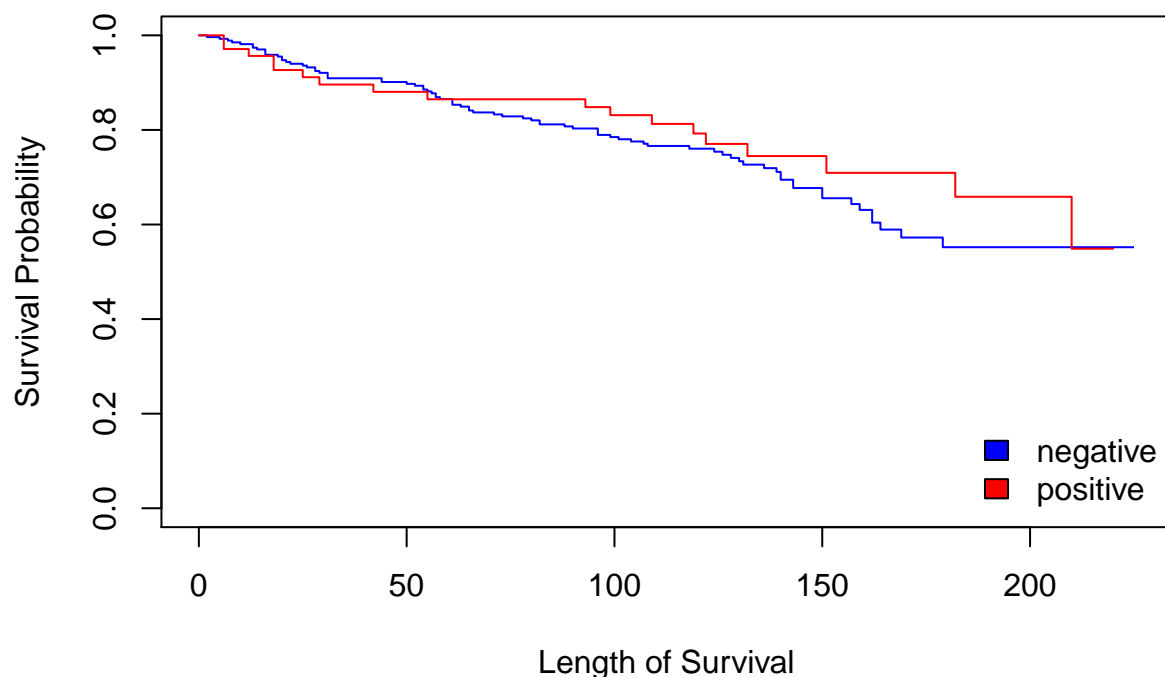
```
plot(surv.HER2,col=c("blue","red"),ylim = c(0,1),
```

```
     main="Plot of Survival Curve by HER2 status",
```

```
     xlab = "Length of Survival",ylab= "Survival Probability")
```

```
legend("bottomright",legend=c("negative","positive"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by HER2 status



```
#levels = 0,1. 0 means negative, 1 means positive
```

```
survdif(Surv(survival_month,survival)~HER2)
```

```
## Call:
```

```
## survdiff(formula = Surv(survival_month, survival) ~ HER2)
```

```
##
```

```
##          N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## HER2=0 271      79    75.7    0.147    0.676
```

```
## HER2=1  70      18    21.3    0.521    0.676
```

```
##
```

```
## Chisq= 0.7  on 1 degrees of freedom, p= 0.4
```

grade The overall grade of the tumor specimen at definitive surgery 0 means low grade, 1 means high grade. Low-grade cancer cells (also known as well-differentiated cancer cells) look more like normal cells and tend to grow and spread more slowly than high-grade cancer cells (poorly differentiated or undifferentiated cancer cells.).

```
surv.grade<-survfit(Surv(survival_month,survival)~grade)
```

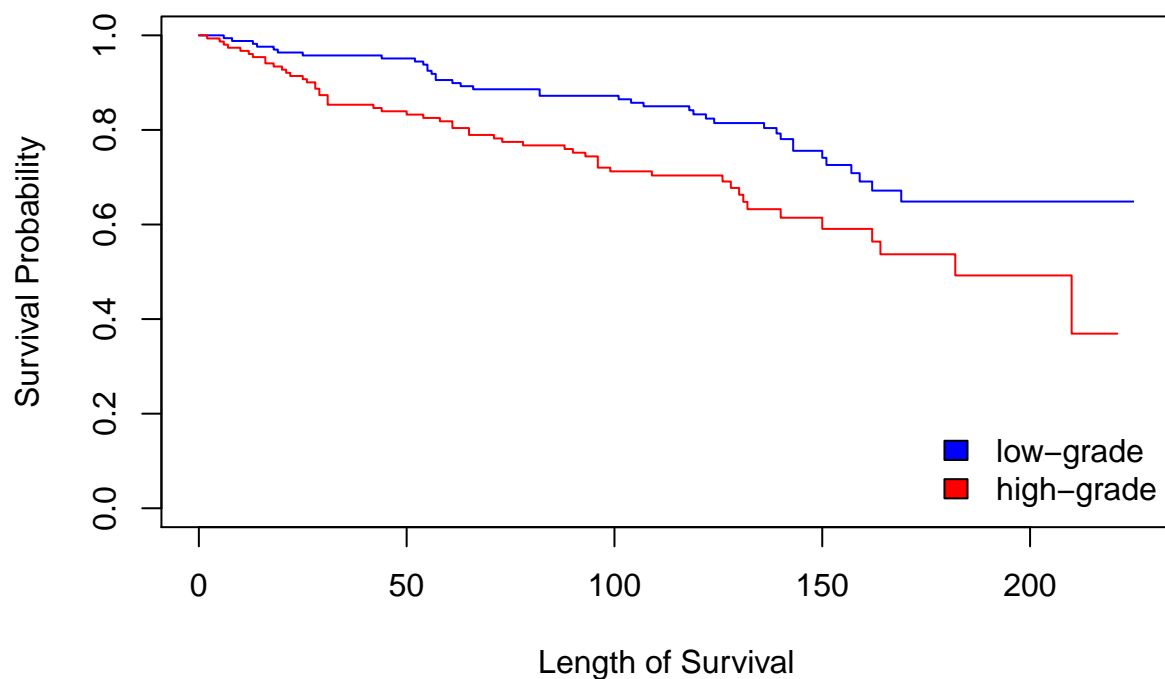
```
plot(surv.grade,col=c("blue","red"),ylim = c(0,1),
```

```
      main="Plot of Survival Curve by Grade Group",
```

```
      xlab = "Length of Survival",ylab= "Survival Probability")
```

```
legend("bottomright",legend=c("low-grade","high-grade"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by Grade Group



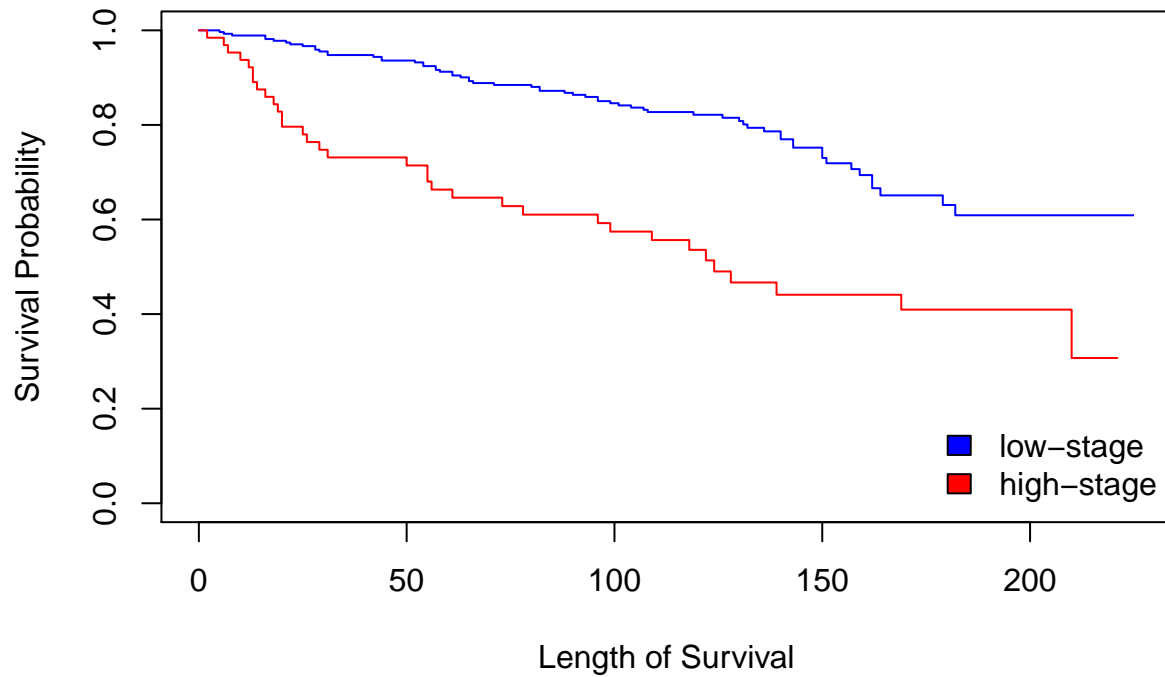
```
survdif(Surv(survival_month,survival)~grade)
```

```
## Call:
## survdif(formula = Surv(survival_month, survival) ~ grade)
##
## n=323, 18 observations deleted due to missingness.
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## grade=0 169      38      52      3.79      8.95
## grade=1 154      53      39      5.06      8.95
##
## Chisq= 9  on 1 degrees of freedom, p= 0.003
```

stage 0 means low stage, 1 means high stage. Lower grade cancers are typically less aggressive and have a better prognosis. The more abnormal the cells look and organize themselves, the higher the cancer's grade. Cancer cells with a high grades tend to be more aggressive.

```
surv.stage<-survfit(Surv(survival_month,survival)~stage)
plot(surv.stage,col=c("blue","red"),ylim = c(0,1),
     main="Plot of Survival Curve by Stage Group",
     xlab = "Length of Survival",ylab= "Survival Probability")
legend("bottomright",legend=c("low-stage","high-stage"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by Stage Group



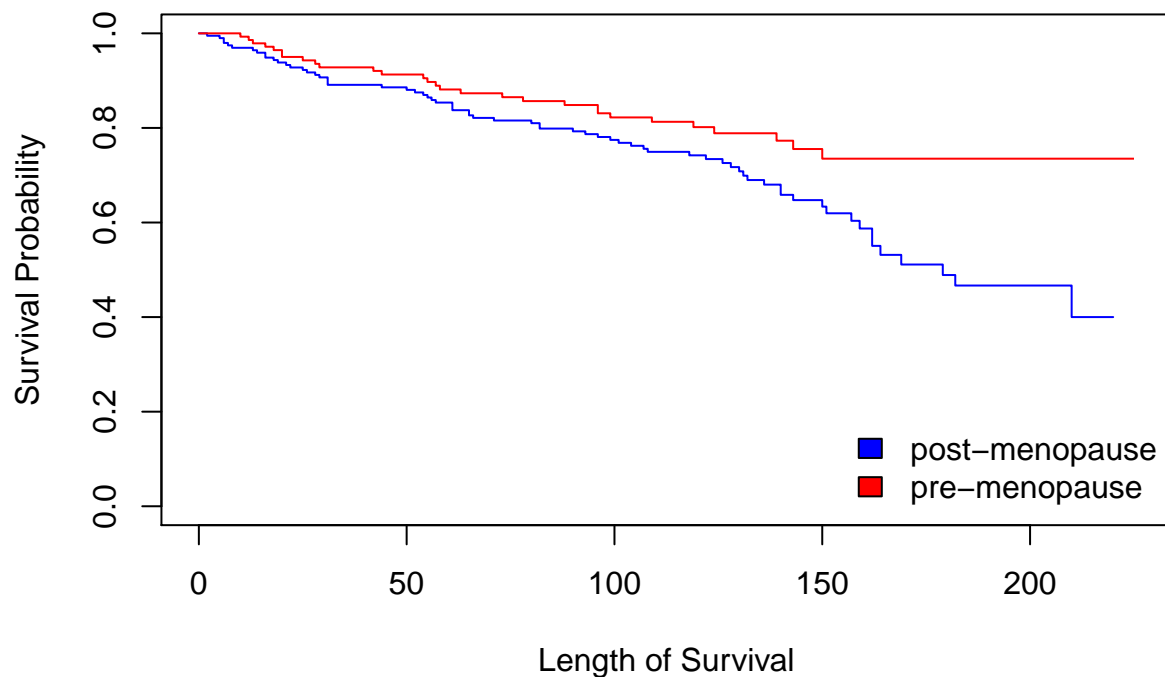
```
survdif(Surv(survival_month,survival)~stage)
```

```
## Call:
## survdif(formula = Surv(survival_month, survival) ~ stage)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## stage=0 277      63    81.1      4.02      24.7
## stage=1  64      34    15.9     20.45      24.7
##
## Chisq= 24.7  on 1 degrees of freedom, p= 7e-07
```

menopause subject's menopausal status at diagnosis 0 means post-menopause, 1 means pre-menopause

```
surv.menopause<-survfit(Surv(survival_month,survival)~menopause)
plot(surv.menopause,col=c("blue","red"),ylim = c(0,1),
     main="Plot of Survival Curve by Menopause Status",
     xlab = "Length of Survival",ylab= "Survival Probability")
legend("bottomright",legend=c("post-menopause","pre-menopause"),fill=c("blue","red"),bty="n")
```

Plot of Survival Curve by Menopause Status



```
survdif(Surv(survival_month,survival)~menopause)
```

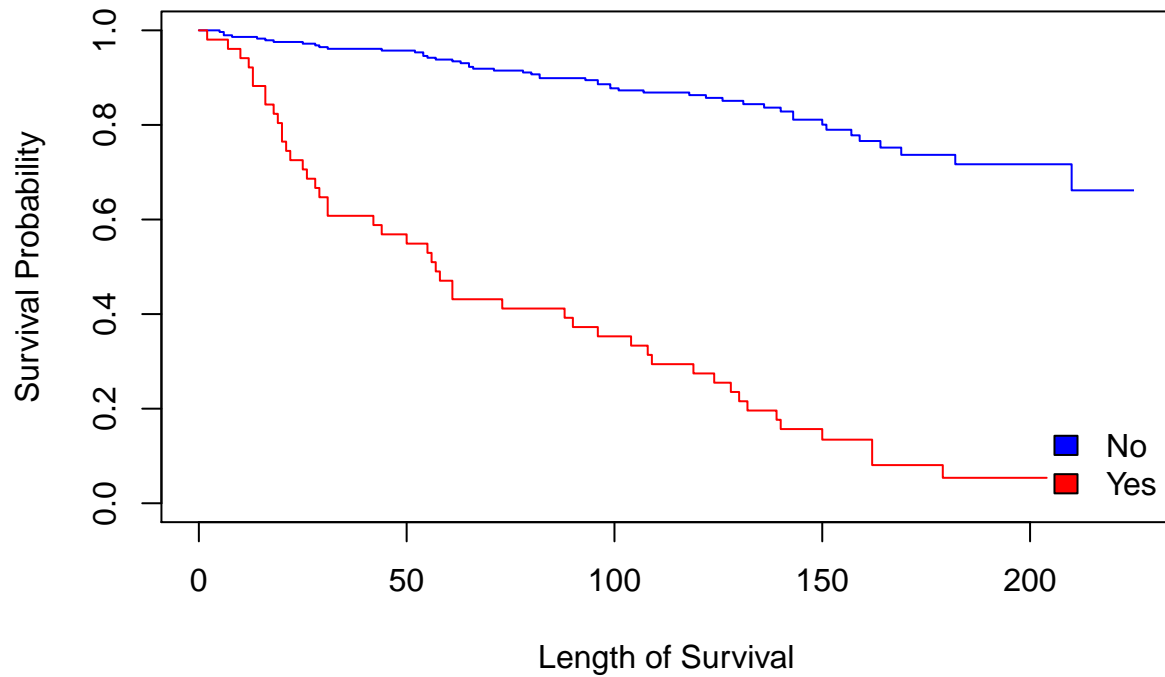
```
## Call:
## survdiff(formula = Surv(survival_month, survival) ~ menopause)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## menopause=0 198      68    55.5      2.79      6.56
## menopause=1 143      29    41.5      3.74      6.56
##
## Chisq= 6.6  on 1 degrees of freedom, p= 0.01
```

This plot looks quite similar to the plot of survival curve by age, Since all old patients are most-menopause and most of young patients are pre-menopause. However, the p-value is not significant.

metastatic Has the subject been diagnosed with metastatic/distant disease?

```
surv.metastatic<-survfit(Surv(survival_month,survival)~metastatic)
plot(surv.metastatic,col=c("blue","red"),ylim = c(0,1),
     main="Plot of Survival Curve by Metastatic Status",
     xlab = "Length of Survival",ylab= "Survival Probability")
legend("bottomright",legend=c("No","Yes"),fill=c("blue","red"),bty="n")
```


Plot of Survival Curve by Metastatic Status



```
survdif(Surv(survival_month,survival)~metastatic)
```

```
## Call:
## survdiff(formula = Surv(survival_month, survival) ~ metastatic)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## metastatic=No 290      50    87.4      16      163
## metastatic=Yes  51      47     9.6     146      163
##
## Chisq= 163 on 1 degrees of freedom, p= <2e-16
```

```
table(grade,metastatic)
```

```
##      metastatic
## grade  No  Yes
##    0 155  14
##    1 120  34
```

```
table(survival,metastatic)
```

```
##      metastatic
## survival  No  Yes
##    0 240   4
##    1  50  47
```

The metastatic feature is dominant in survival. There are 51 patients who were diagnosed with metastatic/distant disease and only 4 of them survived at the censored time.

Include age, PR, stage, in the report results section.