Homework 7

Snigdha Peddi

References

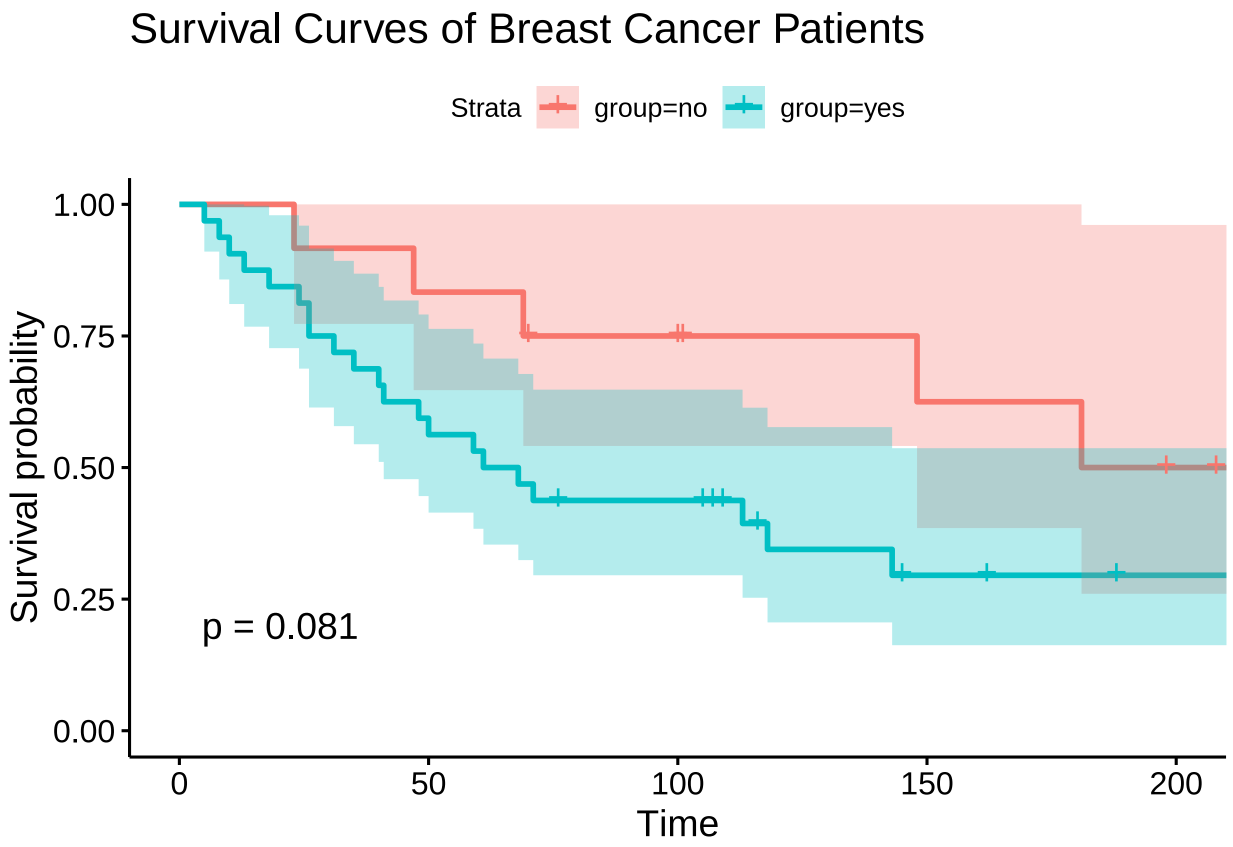
- Lecture code
- Ani Katchova, **Survival Analysis**, 2013,(https://sites.google.com/site/econometricsacademy/econometrics-models/survival-analysis)
- Blogpost by Alboukadel Kassambara,Marcin Kosinski,Prcemyslaw Biecek, **survminer:Survival Analysis and Visualization**,(https://rpkgs.datanovia.com/survminer/)
- Lecture by Jonatan Lindh,**Logrank test in R**,OCtober 14,2016,(https://www.youtube.com/watch?v=HvrBFRzuCvA)

Exercises

Question 1. (Question 11.2 on pg. 224 in HSAUR, modified for clarify) A healthcare group has asked you to analyze the **mastectomy** data from the **HSAUR3** package, which is the survival times (in months) after a mastectomy of women with breast cancer. The cancers are classified as having metastasized or not based on a histochemical marker. The healthcare group requests that your report should not be longer than one page, and must only consist of one plot, one table, and one paragraph. Make sure to keep track of the assumptions that go into a Kaplan-Meier test. Be explicit about what you are actually testing (hint: What types of censoring allows you to still do a valid test?)

a. Plot the survivor functions of each group only using ggplot, estimated using the Kaplan-Meier estimate. **b.** Use a log-rank test (using `logrank_test()`) to compare the survival experience of each group more formally. Only present a formal table of your results. **c.** Write one paragraph summarizing your findings and conclusions.

Plot of Kaplan-Meier Estimate



Log-rank test

```
##   pvalue.of.Logrank_test
## 1             0.06146077
```

Summary

Reviewing the survival object show that the event has occurred for 26 patients (death) and 18 patients are censored. The mean time of the patients survival after the mastectomy is 46.33 months.The Kaplan-Meier estimator was fit and shows that 26 events have occurred. The summary of the fit shows that for the group, metastasized the last event observed was at 143 months and has a survival probability of 0.295 and for the other group not based on histochemical marker,the last event occurred at 181 months and has a survival probability of 0.5.The Survival probabilities over time of both groups can be observed in the plot of the fit.The logrank_test is a non-parametric test that allows to estimate the survival function and can be used to compare the survival curves of 2 groups. The p value is marginal and not significant stating that the null hypothesis that the two groups do not differ in terms of survival is accepted.From the plot the probability of survival of patients classified as metastasized is less compared to the other group but as the p value suggests both groups are not significantly different.The p value obtained from the ggsurvplot function is 0.081 and from the Z statistics of logrank test is 0.06146077. Both the values suggest same thing that both the groups are not different.

Question 2. An investigator collected data on survival of patients with lung cancer at Mayo Clinic. Use the **cancer** data located in the **survival** package. Write up in a narrative style appropriate for the statistical methods section of a research paper/technical report, making sure to address the following points of interest. Use a writing style appropriate for your field of work. Submissions that are not a formal write-up will receive zero credit for this portion of the assignment.

a. What is the probability that someone will survive past 300 days?
b. Provide a graph, including 95% confidence limits, of the Kaplan-Meier estimate of the entire study.
c. Is there a difference in the survival rates between males and females? Make sure to provide a formal statistical test with a p-value and visual evidence.
d. Is there a difference in the survival rates for the older half of the group versus the younger half? Make sure to provide a formal statistical test with a p-value and visual evidence.

OVERVIEW:

The Survival data of patients with advanced Lung Cancer is acquired from the North Central Cancer Treatment group and is stored as **cancer** data in the **survival** package.I will be computing the estimate of survival curve for the data using Kaplan-Meier estimate and study the survival probabilities of the Lung cancer patients at a perticular time during study,review the differences in the survival rates of the patients at different levels of covariates, age and sex.Discuss the statistical significance of these levels using logrank test with supporting plots.

INTRODUCTION:

The Cancer data has records of 228 patients and has 12 variables. The covariates are institution code, Survival time in days, Status where the patient is dead or censored by end of the study, Age in years, Sex, ECOG performance scores, Karnofsky performance score related by physician and by patient, Calories consumed at meals and Weight loss in last six months. For the survival analysis the survival time in days and the status variables were used.

ANALYSIS AND DISCUSSION:

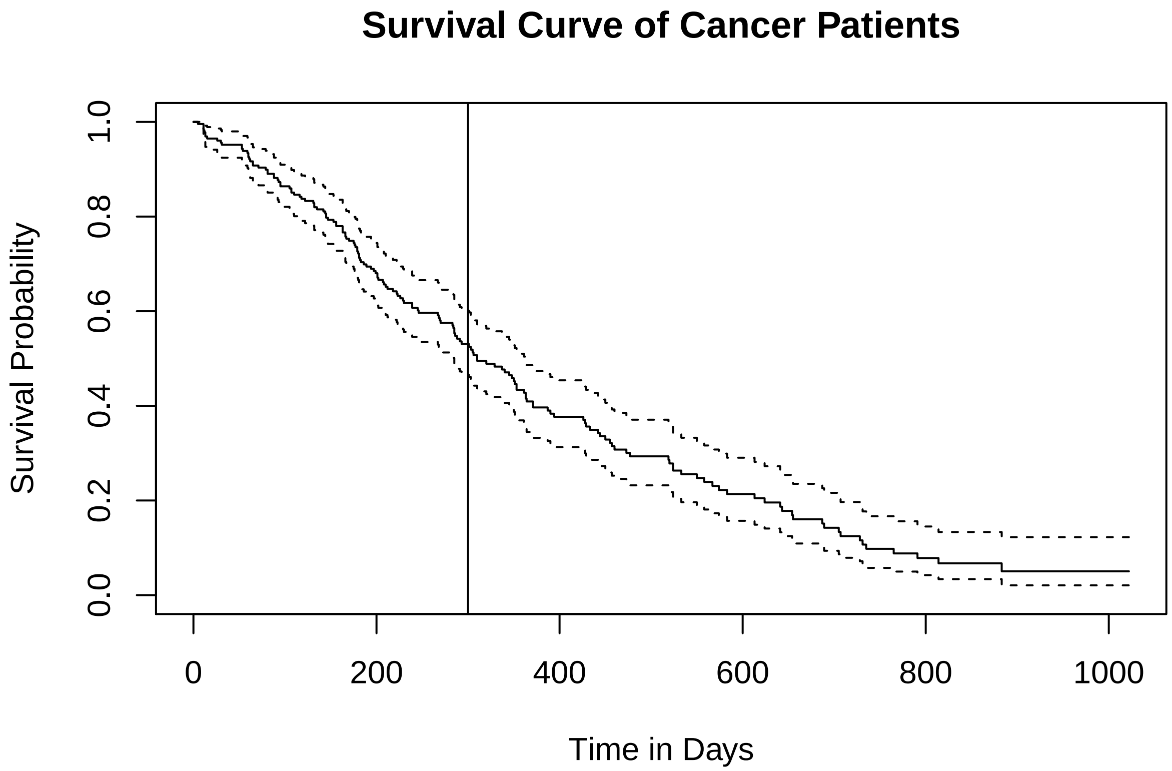
A Kaplan-Meier estimate is fit for the cancer data where time in days and the event(dead) is used as the dependent variable for survival curve.A *Surv* object is created and computed by the function *survfit*.

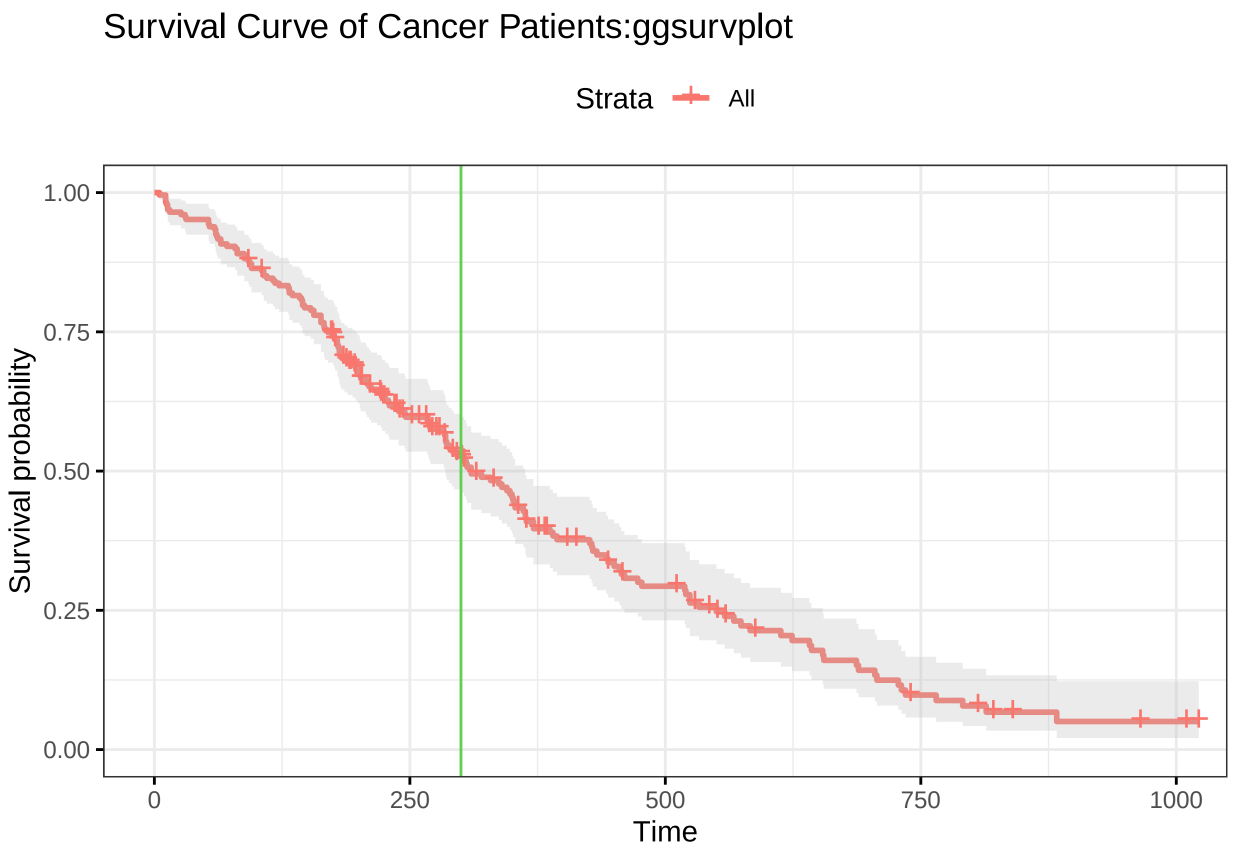
```
  surv <- Surv(time1,event1)
  cancerfit1 <- survfit(surv~1,data=cancer)
```

The summary of the estimate gives the description of the fit with time,number of patients at risk at given time,number of events occurred,survival probabilities at that time,lower and upper confidence intervals. The probability of someone surviving past certain time, probability of surviving to the end of the study, probability of survival past the study,number of events occurred ,number of patients at risk and other survival problems can be solved from the summary of the fit. Below is the survival probability of patients past 300 days.

```
## The probability that someone will survive past 300 days:  0.5306081
```

The summary of the fit can be clearly understood from the plot.Additionally, *ggsurvplot* function from **survminer** package is used to create a similar plot.A vertical reference line in both the plots at survival time of 300 days, further confirms the survival probability of someone past 300 days to be around 0.53.


Survival Curve of Cancer Patients
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Time in Days



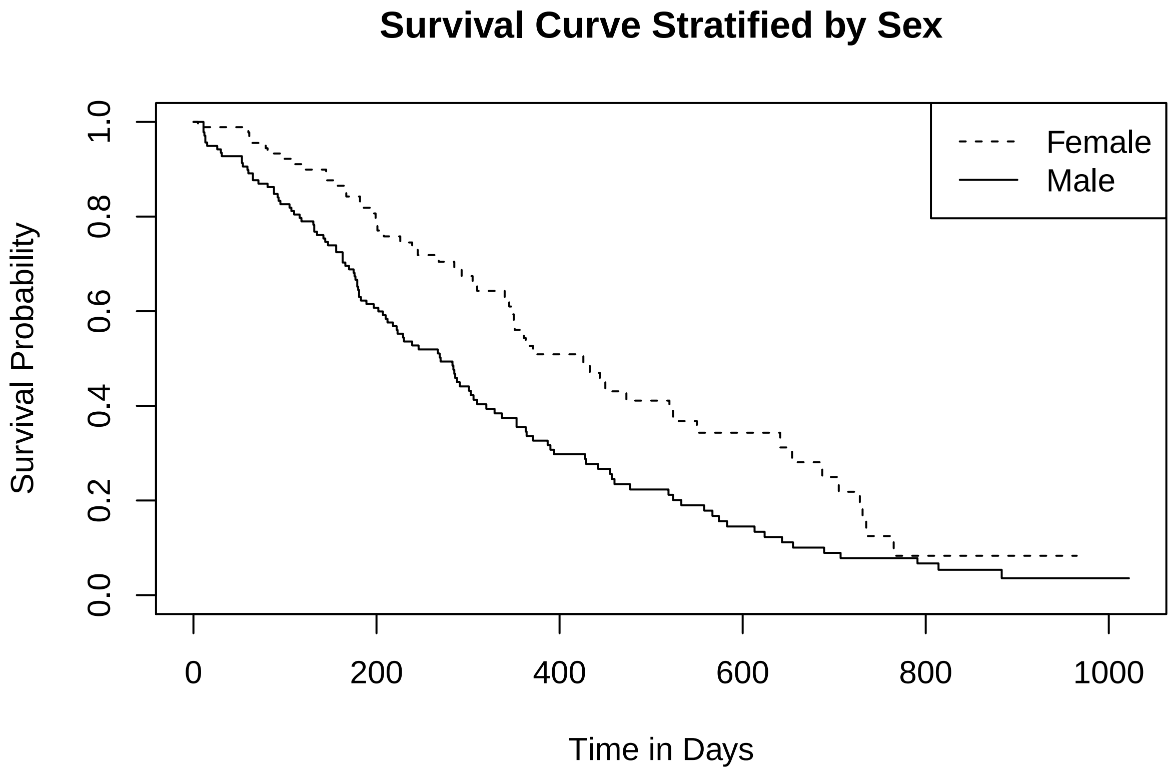
The difference in the survival rates between males and females is studied. The different levels in the sex variable are Male and Female. A Kaplan-Meier estimate is fit by group for sex variable.

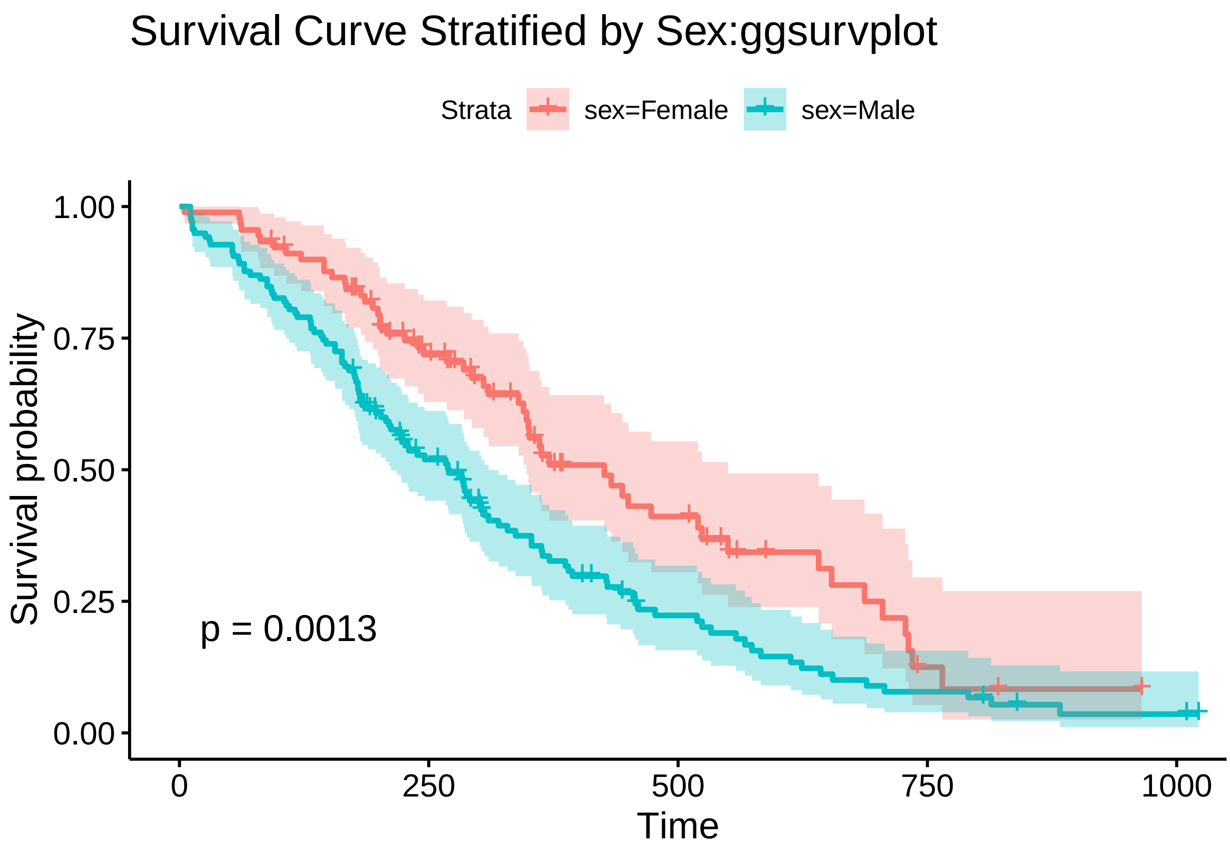
```
cancerfit.sex <- survfit(Surv(time1,event1)~sex,data=cancer)
```

A logrank test is performed to check if survival rates differ between males and females.

```
## p value of survival rates by sex:  0.001
```

A p value of 0.001 from Z statistics shows that the survival rates are significantly different between Male and Female patients at 95% confidence interval and I reject the null hypothesis that both the groups do not differ in terms of survival. This is further confirmed from the survival curve plots. The survival rate of the male patients is lower than the female patients.At 95% confidence interval both the curves are separated and the difference is significant. The p value obtained from *ggsurvplot* function is same as the p value obtained from Z statistics of logrank test and indicate that both the groups are significantly different.


Survival Curve Stratified by Sex
Female
Male
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Time in Days



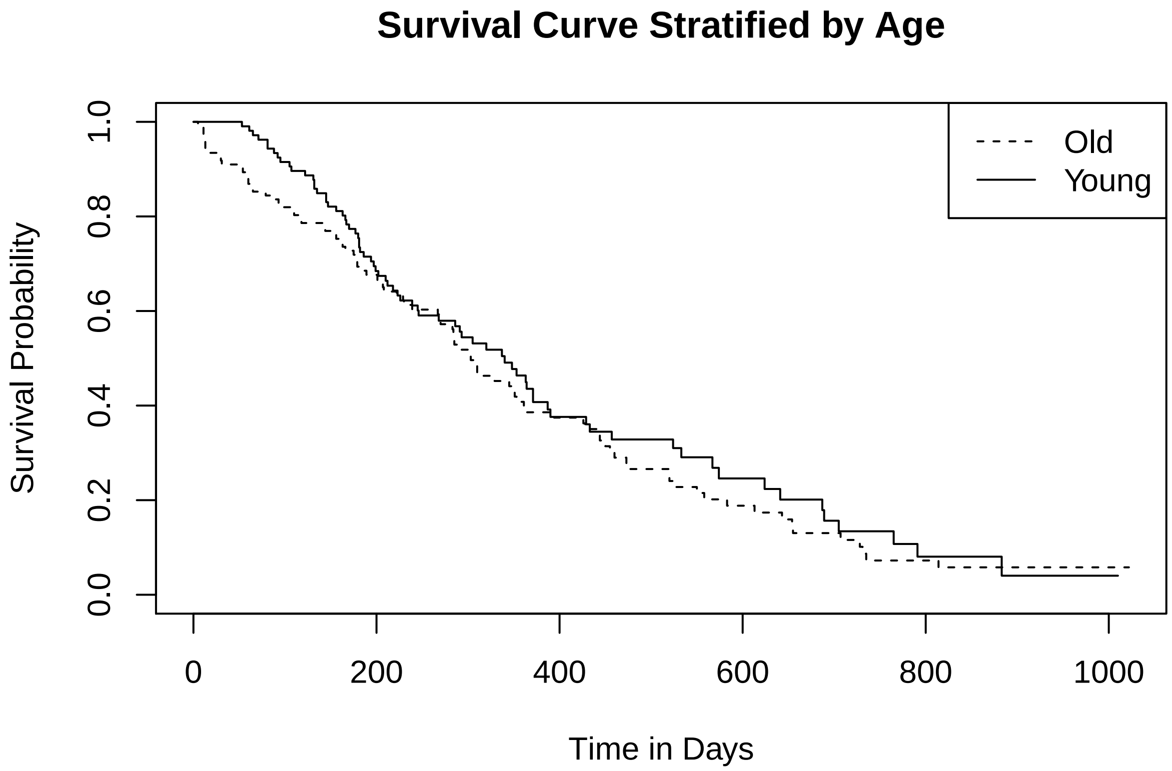
The difference in the survival rates between older half of the group and younger half of the group is studied. A median value of 63 years is used to divide the patients into two groups. All the patients below 63 are considered Young and above 63 are considered as Old.A Kaplan-Meier estimate is fit by group for age variable.

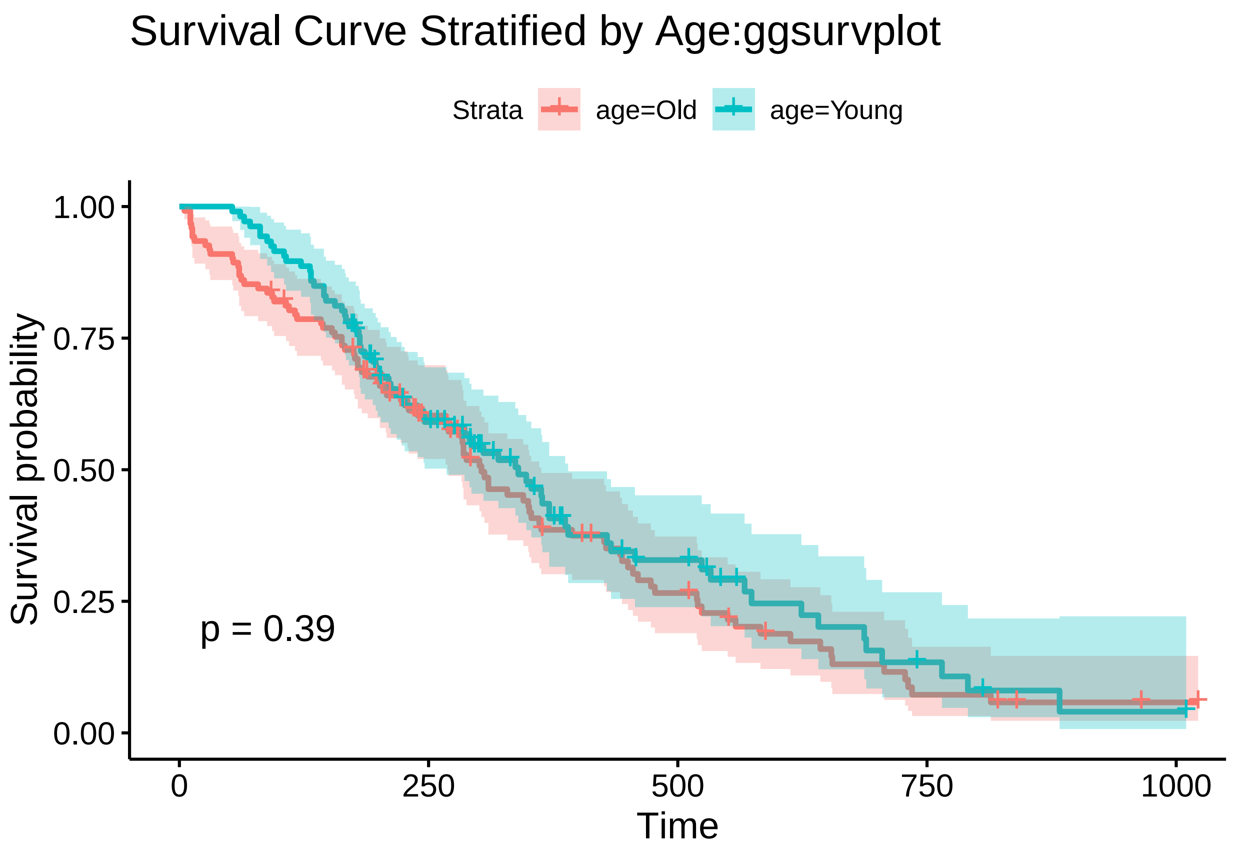
```
cancerfit.age <- survfit(Surv(time1,event1)~age,data=cancer)
```

A logrank test is performed to check if survival rates differ between Younger and Older Patients.

```
## p value of survival rates by age:  0.388
```

A p value of 0.388 from Z statistics shows that the survival rates are not significantly different between Younger and Older patients at 95% confidence interval and we accept the null hypothesis that both the groups do not differ in terms of survival. This is further confirmed from the survival curve plots. The survival rate of the Older patients is slightly lower than the Younger patients.At 95% confidence interval both the curves are overlapped and the difference is not significant. The p value obtained from *ggsurvplot* function is same as the p value obtained from Z statistics of logrank test and indicate that both the groups are not significantly different.


Survival Curve Stratified by Age
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Time in Days
Old
Young



CONCLUSION:

The Survival data of patients with advanced Lung Cancer (cancer data) from **survival** package is analyzed and Kaplan-Meier estimates were computed for survival curves.The survival probabilities of patients past 300 days was found to be 0.53. Survival curves were fit to investigate if the gender and age effects the survival probability.Logrank test is performed to obtain the p values.A significant p value of 0.001 from the fit by sex indicate that the the survival probabilities are significantly different between Male and Female patients at 95% confidence interval and from the plot it is clearly visible that the Female patients have higher survival probabilities than Male patients. Also, a p value of 0.39 from the fit by age (<63-Young, >63-Old) indicate that the age do not have significant effect on the survival probabilities and not significant at 95% confidence interval.The Survival curve confirms this analysis where both the curves of Young and Old patients are very close.