Survival Analysis From HSAUR

Yamuna Dhungana

Exercises

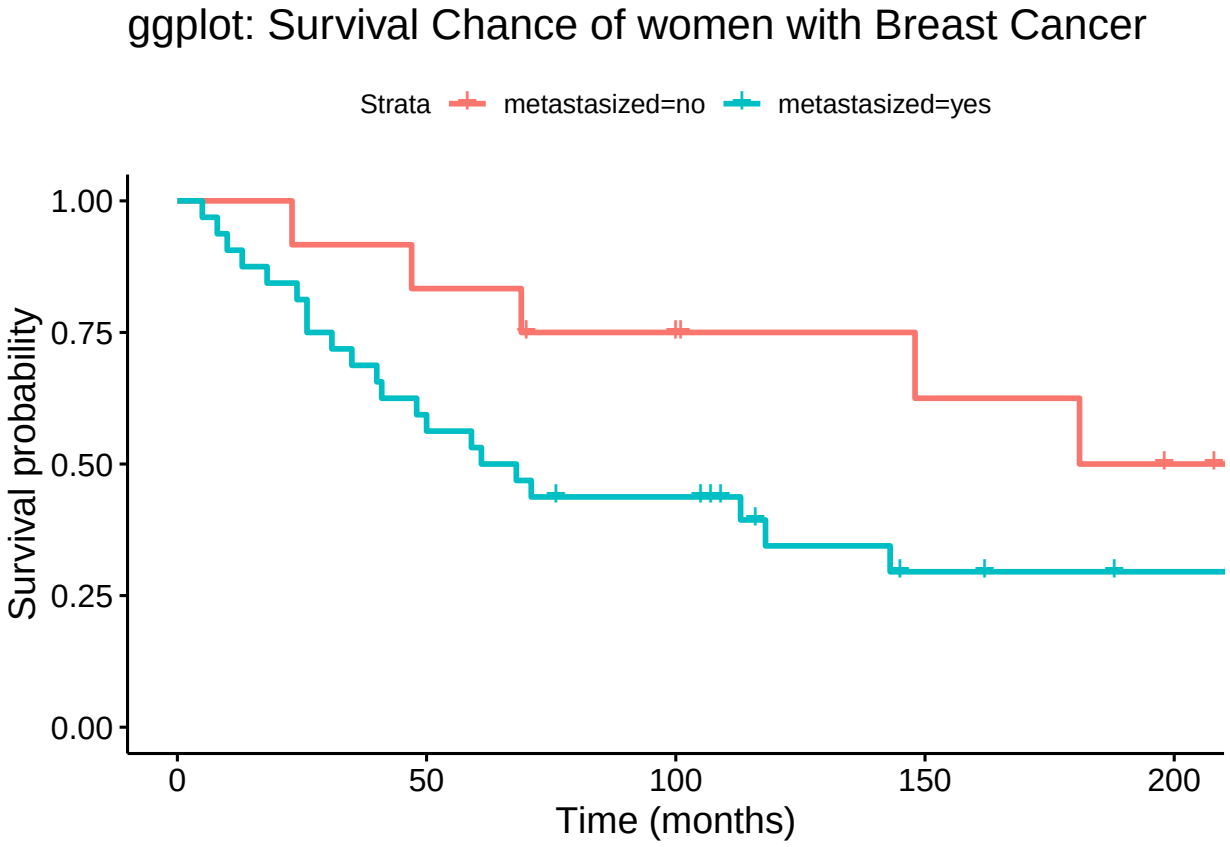
1. (Question 11.2 on pg. 224 in HSAUR, modified for clarify) A healthcare group has asked you to analyze the **mastectomy** data from the **HSAUR3** package, which is the survival times (in months) after a mastectomy of women with breast cancer. The cancers are classified as having metastasized or not 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.

```
## Call: survfit(formula = Surv(time, event) ~ metastasized, data = bcancer_data)
##
##                   n events median 0.95LCL 0.95UCL
## metastasized=no  12      5  181.0     148      NA
## metastasized=yes 32     21   64.5      41      NA
```

```
## Call: survfit(formula = Surv(time, event) ~ metastasized, data = bcancer_data)
##
##                 metastasized=no
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    23     12       1    0.917  0.0798        0.773        1.000
##    47     11       1    0.833  0.1076        0.647        1.000
##    69     10       1    0.750  0.1250        0.541        1.000
##   148      6       1    0.625  0.1545        0.385        1.000
##   181      5       1    0.500  0.1667        0.260        0.961
##
##                 metastasized=yes
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     5     32       1    0.969  0.0308        0.910        1.000
##     8     31       1    0.938  0.0428        0.857        1.000
##    10     30       1    0.906  0.0515        0.811        1.000
##    13     29       1    0.875  0.0585        0.768        0.997
##    18     28       1    0.844  0.0642        0.727        0.979
##    24     27       1    0.812  0.0690        0.688        0.960
##    26     26       2    0.750  0.0765        0.614        0.916
##    31     24       1    0.719  0.0795        0.579        0.893
##    35     23       1    0.688  0.0819        0.544        0.868
##    40     22       1    0.656  0.0840        0.511        0.843
##    41     21       1    0.625  0.0856        0.478        0.817
##    48     20       1    0.594  0.0868        0.446        0.791
##    50     19       1    0.562  0.0877        0.414        0.764
##    59     18       1    0.531  0.0882        0.384        0.736
##    61     17       1    0.500  0.0884        0.354        0.707
##    68     16       1    0.469  0.0882        0.324        0.678
```

```
##     71      15       1    0.438  0.0877        0.295        0.648
##    113      10       1    0.394  0.0892        0.253        0.614
##    118       8       1    0.345  0.0906        0.206        0.577
##    143       7       1    0.295  0.0900        0.162        0.537
```



The above plot is the survival plot of the women whose cancer has spread to the other organs and whose cancer has not spread. In the plot, named Survival chance of women with whose cancer has spread in a whole body. The green line denotes the probability of women who have undergone metastasis. The red line is the probability of women who have not undergone metastasis. The X-axis is the survival time in months, and Y-axis is the probability of survival. The plot shows that survival without metastasized (Redline) has a higher probability of surviving. Likewise, the surviving chances of the patient with metastasized (green line) are lower. The plot also shows the probability of surviving is higher than 0.5 for all the patients who are not metastasized. Whereas, for the metastasized patient, the probability of survival drops down to 0.5 from 50 months. In the period of 60 months to 150 months, the survival of women with metastasized drops to 0.25. Whereas, for the patient without metastasized reveals that survival probability is constant from 60 to 150 and drops to 0.65 at 150 months. Hence from the graph, I can say that patient without metastasized has a comparatively lower risk of death than that with metastasized.

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.

```
## Log rank test
```

```
## Call:
## survdiff(formula = Surv(time, event == 1) ~ metastasized, data = bcancer_data,
##     rho = 0)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
## metastasized=no  12        5      9.2      1.91      3.04
## metastasized=yes 32       21     16.8      1.05      3.04
##
##  Chisq= 3  on 1 degrees of freedom, p= 0.08
```

```
## cox regression
```

```
## Call:
## coxph(formula = Surv(time, event) ~ metastasized, data = bcancer_data)
##
##                    coef exp(coef) se(coef)     z    p
## metastasizedyes 0.8516    2.3434   0.5022 1.696 0.09
##
## Likelihood ratio test=3.35  on 1 df, p=0.06704
## n= 44, number of events= 26
```

Table 1: P values using different functions

Logrank_test	survdiff	coxph
0.062	0.081	0.09

I used the Log-rank test from the coin library and survfiff for comparing whether the data is statistically significant. I also used cox regression. From the survdiff and log-rank test, I got the p-value 0.08 and 0.06. which is not statistically significant at $p < 0.05$. Whereas, from the cox-regression and I got the p-value 0.089 which is also not significant at 0.05.

c. Write one paragraph summarizing your findings and conclusions.

The pattern of the survival plot assumed that having metastasis is the main factor for a decreased survival chance of women with breast cancer. We also saw the plot supporting the same. However, we performed the log-rank test for the two groups and ended up getting a p-value that is not statistically significant. We saw that the Redline was higher than the green line in our plot. But, according to the log-rank test and the p-value we accept the Null Hypothesis. Cox regression was also performed, which gave a p-value (0.06). So, with a 5% risk level, the difference between the two groups is not statistically significant and hence, the surviving chance of women with breast cancer is not significantly affected by the fact whether the cancer was metastasized or not.But if we take a higher risk level, suppose of 10%, our conclusion is different (since 0.08<0.1). With a 90% confidence level, we can conclude that the surviving chance of women with breast cancer is significantly affected by whether women have cancer was metastasized or not. If cancer spreads to the other organ, the surviving chance significantly decreases. 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?

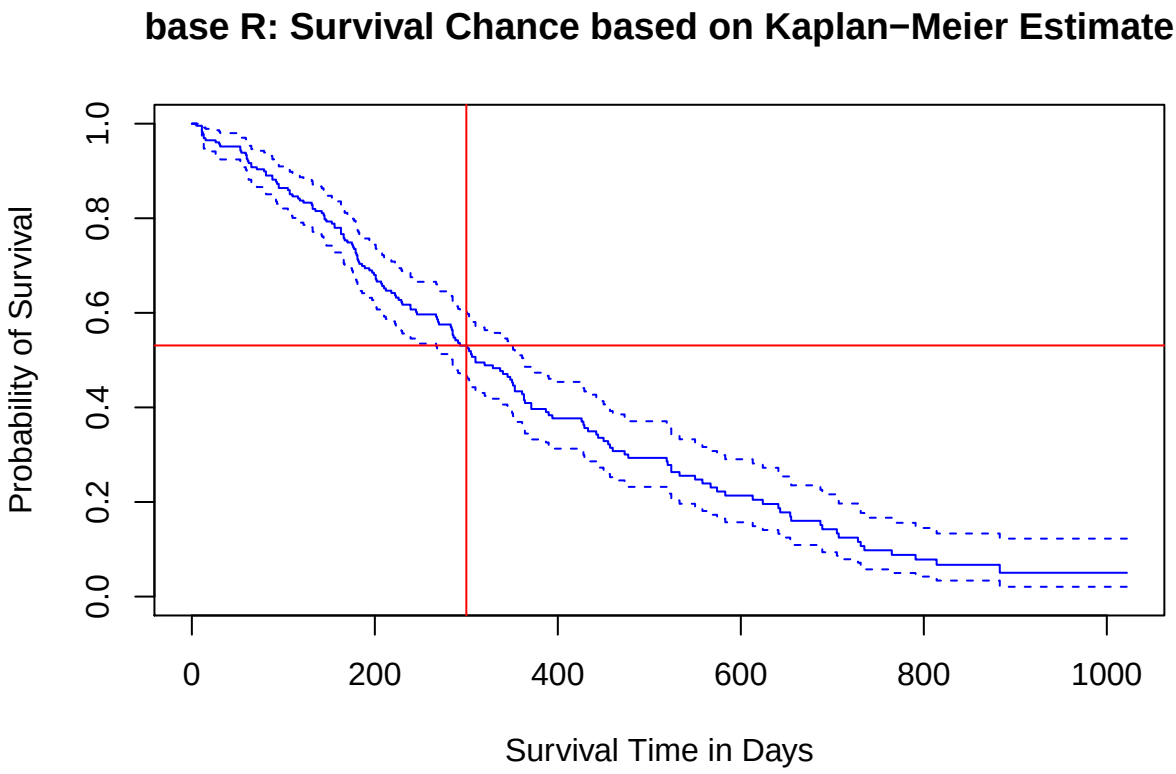
```
## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = cancer)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     92     101    0.531  0.0346        0.467        0.603
```

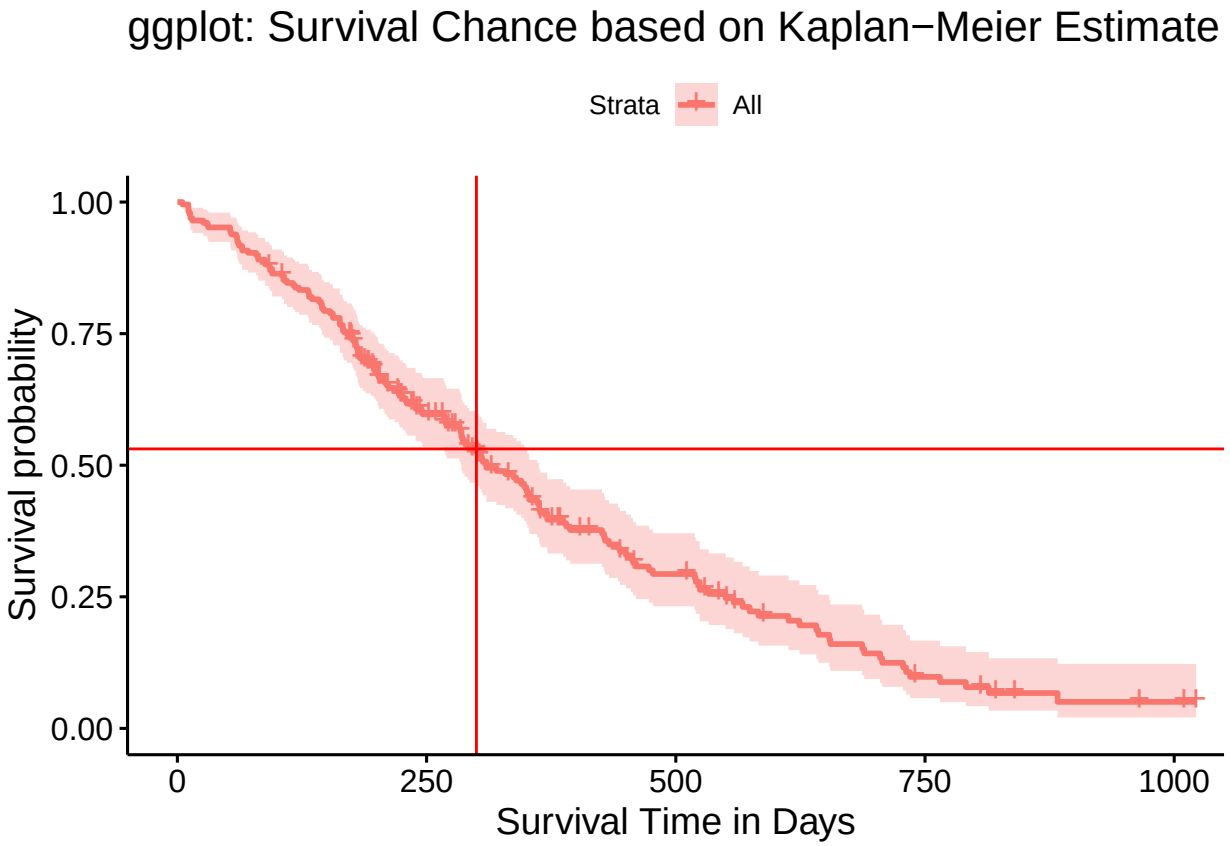
```
## Probability of someone surviving past 300 days
```

```
## $surv
## [1] 0.5306081
```

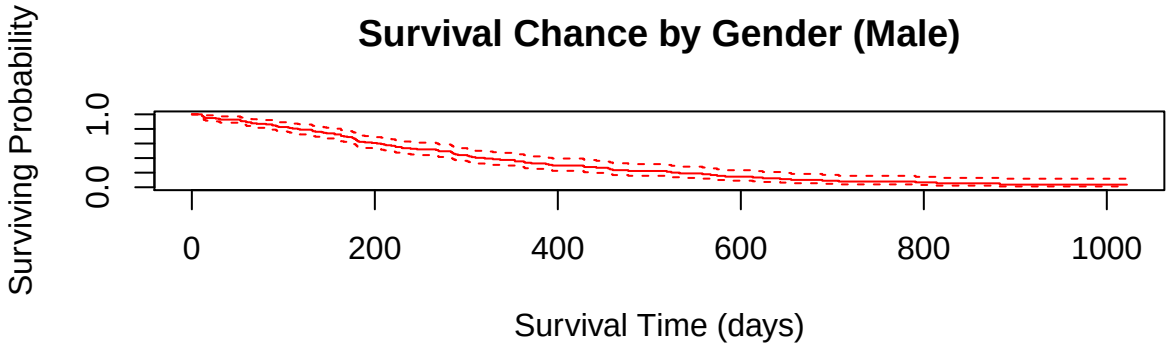
The probability of survival is 0.5306

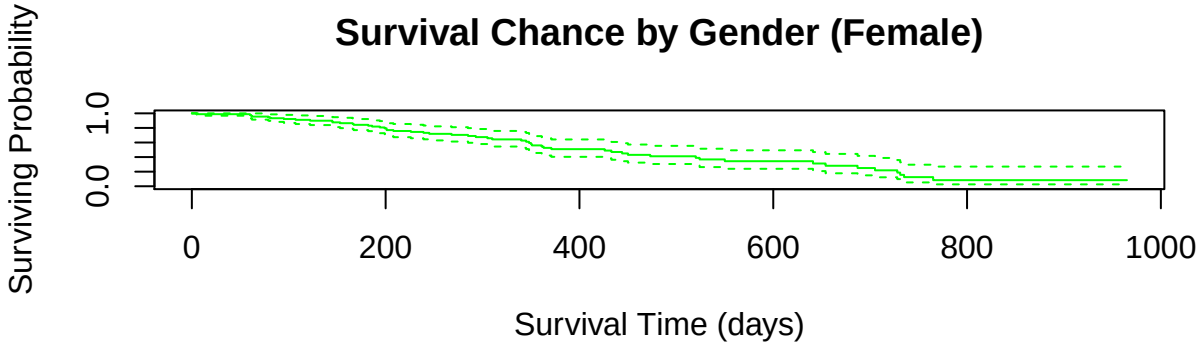
b. Provide a graph, including 95% confidence limits, of the Kaplan-Meier estimate of the entire study.

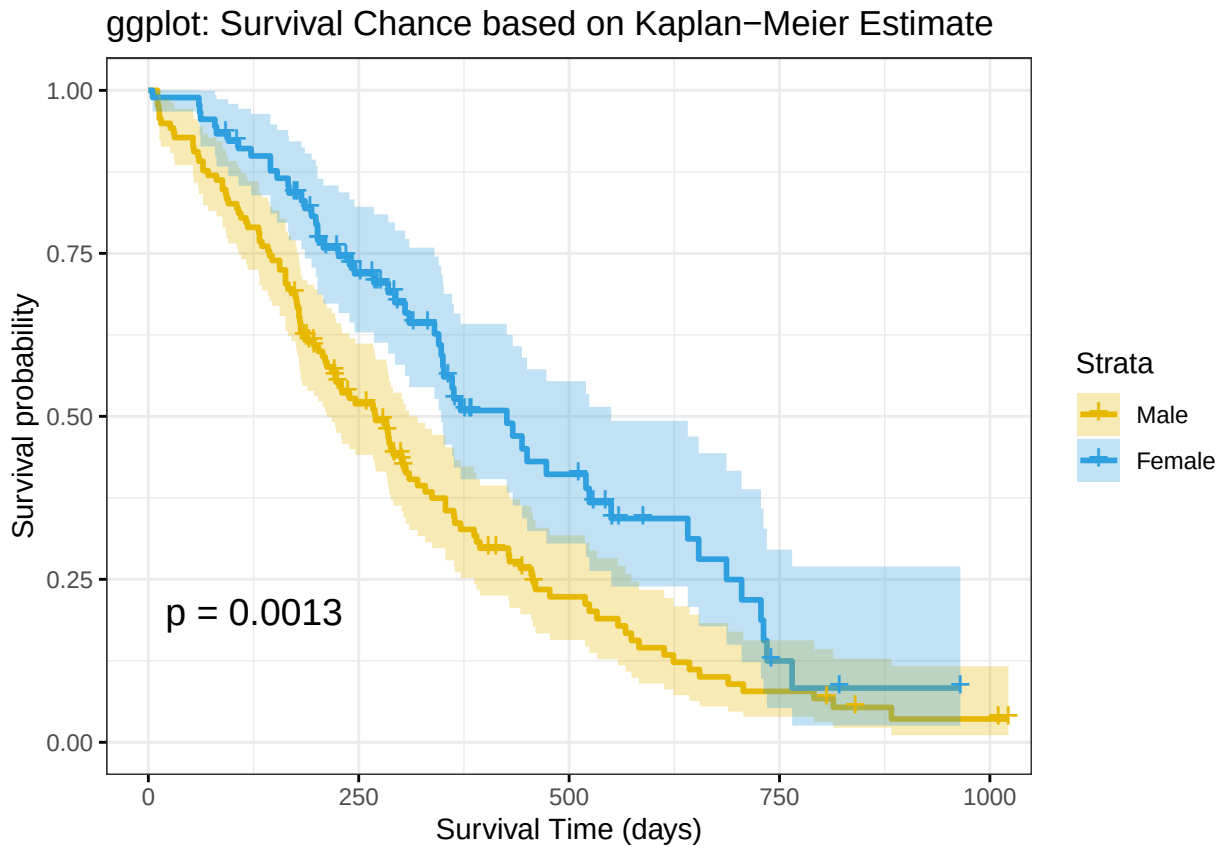




c. Is there a difference in the survival rates between males and females? Make sure to provide a formal statistical test with a p-value and visual evidence.

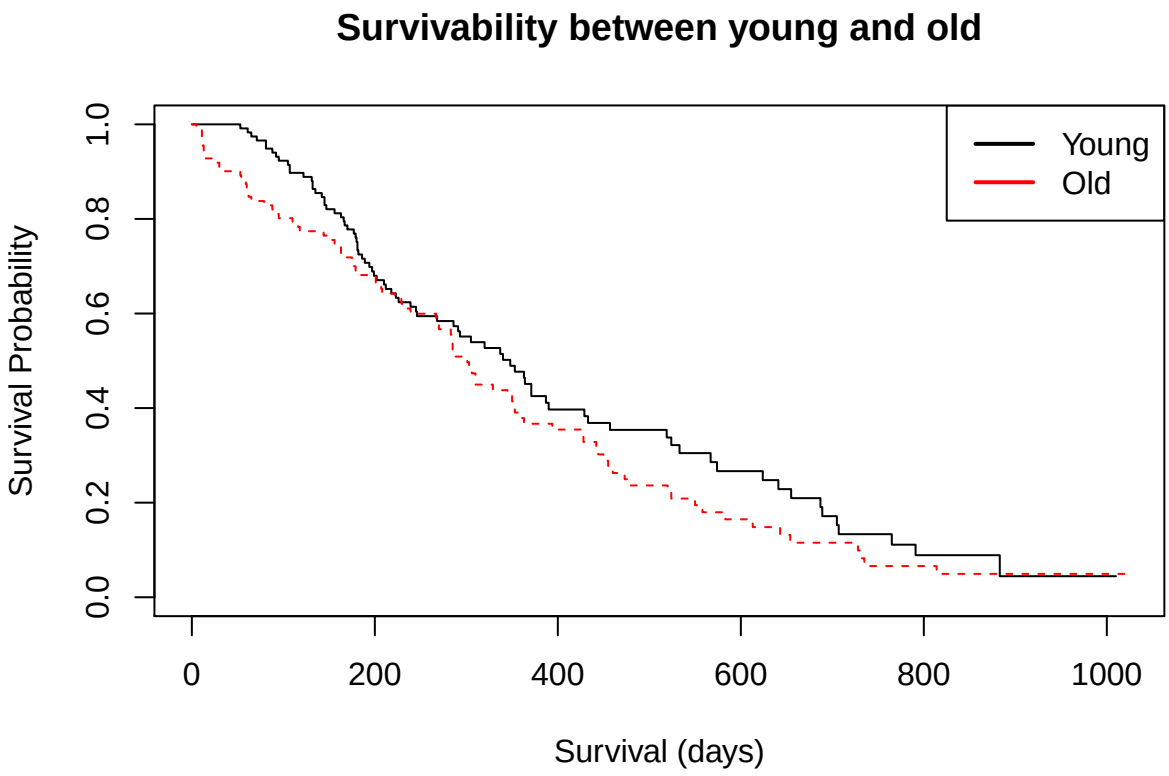

Survival Chance by Gender (Male)
Surviving Probability
0.0
1.0
0
200
400
600
800
1000
Survival Time (days)

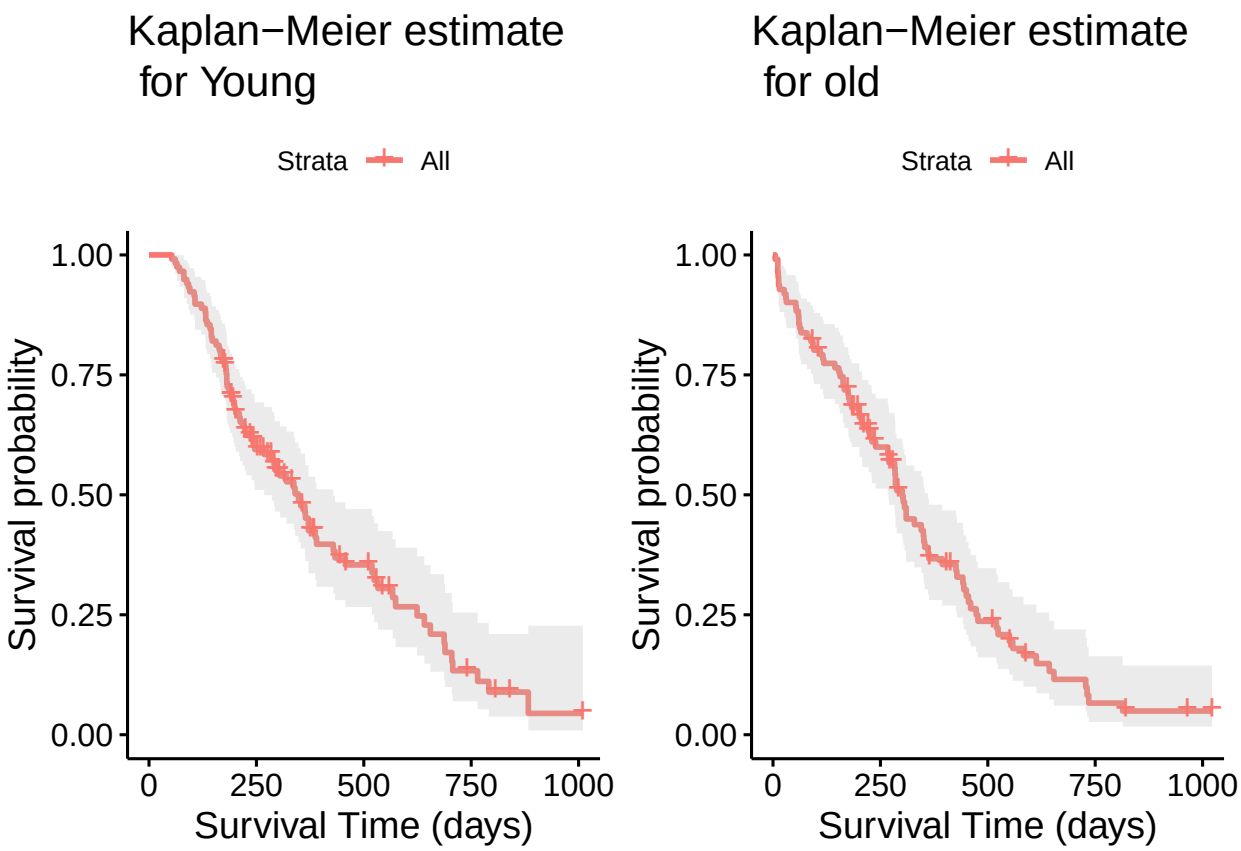

Survival Chance by Gender (Female)
Surviving Probability
0.0
1.0
0
200
400
600
800
1000
Survival Time (days)




```
## Call:
## survdiff(formula = Surv(time, status == 2) ~ sex, data = cancer)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 138      112     91.6      4.55      10.3
## sex=2  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

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.


Survivability between young and old
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Survival (days)
Young
Old




```
## Call:
## survdiff(formula = Surv(time, status == 2) ~ Age_Group, data = c.data)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## Age_Group=1 117       80     88.8     0.865      1.88
## Age_Group=2 111       85     76.2     1.007      1.88
##
##  Chisq= 1.9  on 1 degrees of freedom, p= 0.2
```

Survival of the people with lung cancer, according to age, sex, and past 300 days.

Data collected from a patient with lung cancer at the Mayo Clinic used for analysis. The survival analysis of the people with lung cancer, according to the age group (young and old), and sex, and the past 300 days is determined. To test the survival of the patient, I performed survival analysis. The survival analysis shows gender has a significant influence on the survival of the patient with lung cancer. However, the age group has no significance in a patient with lung cancer. In addition to this, I also performed the probability of surviving the past 300 days with the same analysis. The dataset has ten variables, and we had used only four. Time represents the survival time of the patient in days. Likewise, the status of the patient—1 as censored, and 2 as dead. The parameter sex, where 1 is male, and 2 is female. Dataset was used from the package Survival named cancer.

Firstly, the survival model was fitted (survfit(surv..)) with status 2 means dead. Function survfit gives the probability of 0.5306 and about 53.1%. I have provided the horizontal and vertical lines indicating the probability for the past 300 days. Visual evidence of survival with the confidence of 95%. The plot shows as time increases survival probability also decreases and becomes uncertain past 750 days.

Secondly, for determining if sex has a significant influence on the survival of the patient, we divided the data according to the sex. For visual evidence, we have the plot. The plot shows, males and females have

different survival probability and were determined using Kaplan- Meier estimator.The blue line determines the female's survival probability over time, and the yellow line determines the survival probability of males over time. In females, the survival for a couple of months, in the beginning, is constant and is high. And the survival probability past 750 days is very low and uncertain. Compared to males, survival is high. Likewise, in males, the plot shows the steep without steps, which shows that the risk of male surviving compared to female is low and are at higher risk. Additionally, Log-rank also indicates a significant difference between the two groups. The P-value of 0.001 means that the null hypothesis (There is no difference in male and female) rejected with 95% confidence.

Similarly, by the age group (young and old), a comparison is made. The median-age was determined (by code), and two sub-datasets found Young and old. One means young and two as old. With the Kaplan-Meier estimator, the plot is obtained. The plot, "Survivability between young and old", is the survival of young and old aged patients, black representing young and red representing old. According to the plot, the survivability of the young is higher than the old. However, based on the log-rank, the difference is not statistically significant because of a p-value 0.2. and 0.2 is much higher than 0.05. So, with the 95% confidence interval, we accept the null hypothesis, and we conclude that the age group has no significant difference in survivability. Both age groups are at the same risk.

Hence, from my analysis, the survival of the patient with lung cancer has a significant influence on sex but not by the age group.