Homework 7

Amin Baabol

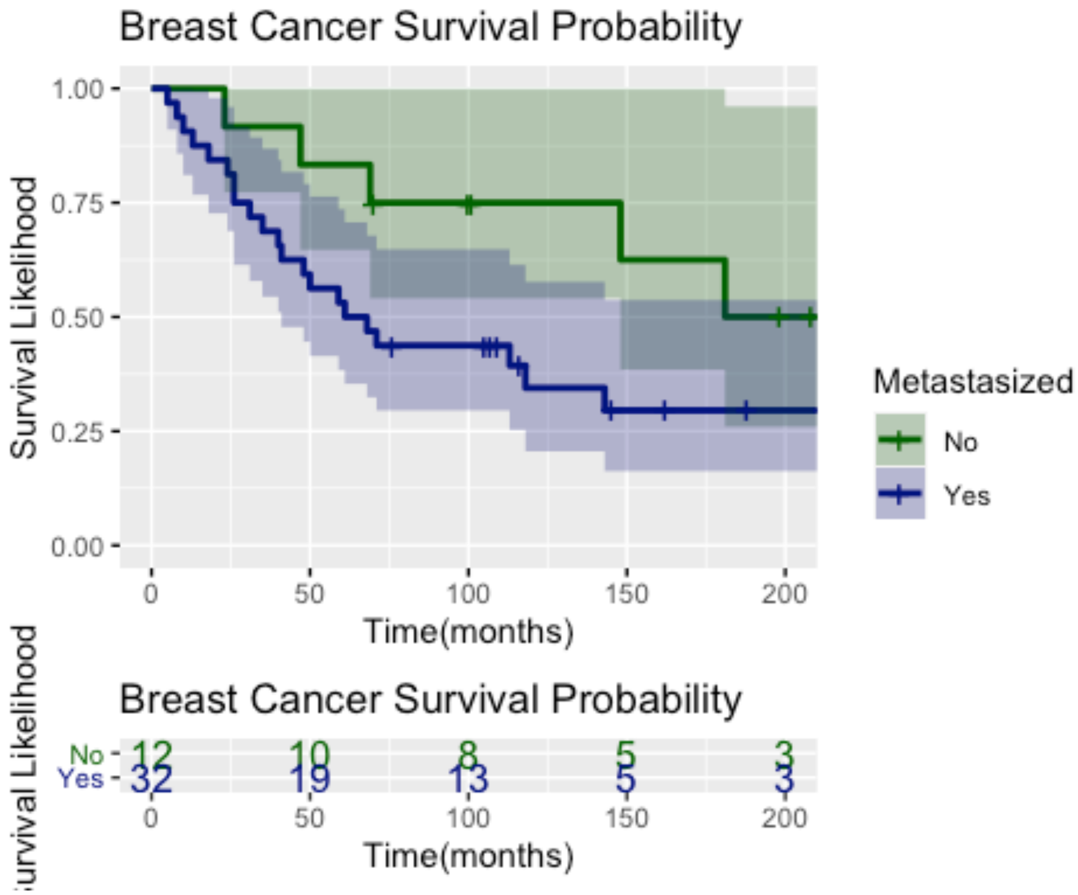
Note: Mohamed and I collaborated in producing this report. We worked together and verified each other's work on every part of this assignment.

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 == 1) ~ metastasized, data =
mastectomy)
##
##                 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
```



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 Statistical Significance 0.08129829
```

c. Write one paragraph summarizing your findings and conclusions.

Summary

We begun the breast cancer survival analysis by fitting the mastectomy data into a Kaplan-Meier model.The plot compares the survival likelihood probability of breast-cancer patients with metastasis against fellow breast-cancer patients but without metastasis.The legend on the right describes with color schemes the two aforementioned groups with green representing non-metastasized breast-cancer patients and blue representing breast-cancer patients have had cancer malignantly spread in other parts of the body.The blue and green shading indicate the lower and upper bounds for the 95% confidence interval.Taking a closer look at plot we can see that the there is an overlay or cross over regions in some of the shaded areas. Looking solely at this plot we would think that the non-metastasized patients have higher probability than their metastasized counterparts, however, after running the logrank test, that deduction doesn't hold up. According to the results of the logrank test, the p-value is 0.08129829 which is greater than our alpha significant level of 0.05. The null hypothesis of this logrank test is that there isn't a difference between the metastasized and the non-metastasized cancer patients in terms of survival likelihood.The alternative hypothesis is that there is a difference in survival probability between the two groups. Given the relatively high p-value of 0.08129829, we fail to reject the null hypothesis at the at the 0.05 significance level.

Title

Understanding Survival Likelihood - Analyzing Mayo Clinic Lung-Cancer Patients Data:

Introduction

The goal of this assignment is to understand the various factors that influence the survival rate or survival probability of lung-cancer patients.In order to address this goal we analyzed data that has been collected on lung-cancer patients at Mayo Clinic.At the end of our analysis we answered the following questions: 1.What is the probability that someone will survive past 300 days? 2.Is there a difference in the survival rates between males and females? 3.Is there a difference in the survival rates for the older half of the group versus the younger half?

Approach

To effectively begin our analysis process, we imported all relevant packages including the **cancer** data from **survival** package.The data contains information such as each patient's survival time, status, age, and sex. These variables will be used to conduct our analysis. The table below lists all the variables, form the Cancer data set, along with their description.

Symbol	Description
$inst$	Institution code
$Time$	Survival time in days
$Status$	censoring status 1=censored, 2=dead
age	Age in years
sex	Male=1 Female=2
$ph.ecog$	ECOG performance score as rated by the physician. 0=asymptomatic, 1= symptomatic but completely ambulatory, 2= in bed <50% of the day,

3= in bed > 50% of the day but not bedbound, 4 = bedbound | $ph.karno$ | Karnofsky performance score (bad=0-good=100) rated by physician | $pat.karno$ | Karnofsky performance score as rated by patient | $meal.cal$: | Calories consumed at meals | $wt.loss$ | Weight loss in last six months ——————————————————————————

Model

Given the cleaned data, we proceeded to constructed a model utilizing the function 'survfit' in the **survival** package.This function facilitates the use of Kaplan-Meier's method in predicting survival probability over specified time.This method is particularly important in our computations because we're dealing with non-parametric and at the same time discreet data.

To address the first questions, we set up and fitted a model using time and status as the parameters and utilized surffit and surv functions to create the survival object, We then analyzed the rates of occurrence of events over time. Furthermore, the summary of the statistics is printed along with visual aids in the form of base R plots and **ggsurvplot** from **ggplot2** package.Moreover, we set up a second model utilizing the same technique and packages for fitting model and plotting the summary. In order to verify the validity of our model, we set up a Log-Rank test using the survdiff function from the **survminer** package. Due to its non-parametric nature it makes no assumptions as far as the distributions of the data. This makes Log-Rank test exceptionally useful in comparing the survival likelihoods of two or more groups with significant accuracy.

Results

a. What is the probability that someone will survive past 300 days?

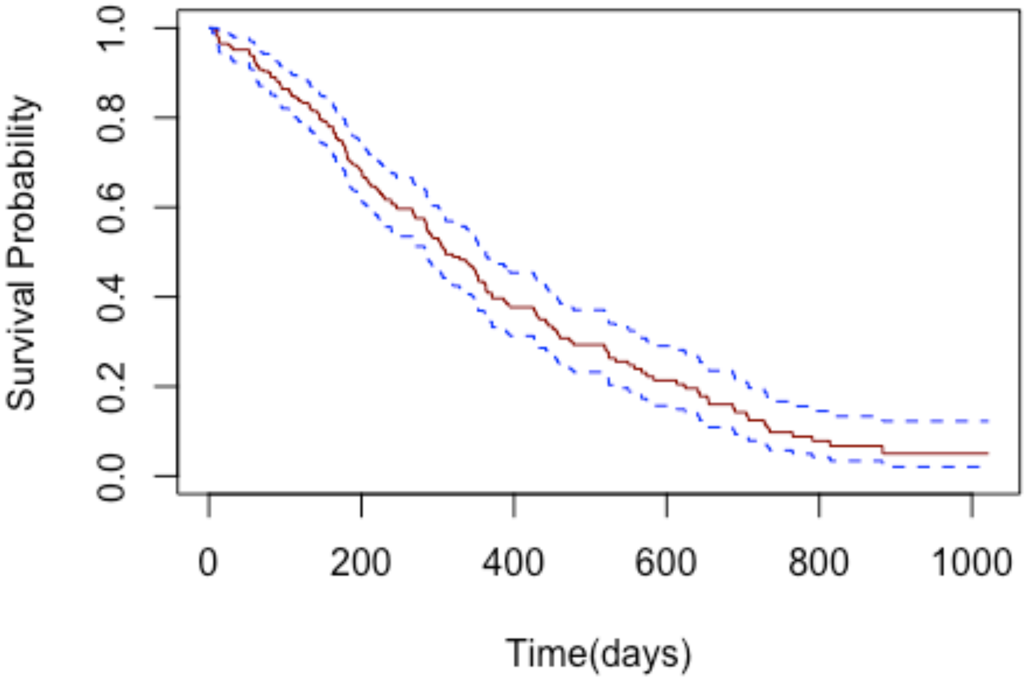
We constructed a model using surv function to estimate the chance of patient surviving past 300 days. We entered the variables time and and status == 2 (2= dead) as parameters for the model. The model estimated that a patient's chances of survival is *0.53* after 300 days.

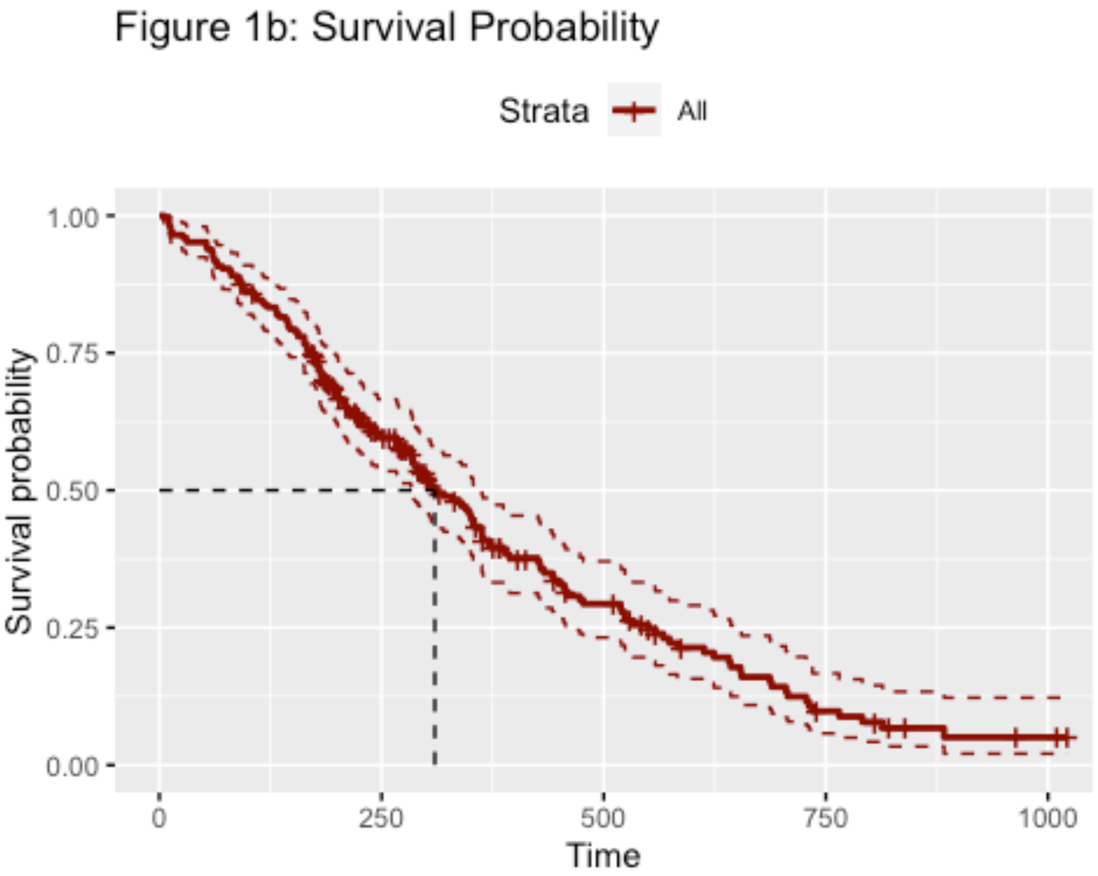
```
## Probability of survival past 300 days is 0.5306081
```

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

The graph show that half of the patients have less than 0.5 chances of survival after 310 days.

Figure 1a: Survival Probability





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.

From the plot, we can see that males have 0.5 and less chance of survival after 270 day of Lung cancer. On the other hand, females have 0.5 and less chance of survival after 426 days. We can conclude that females with lung cancer have higher chances of survival over time than males. To verify out conclusion, we used Log-Rank test to see if there any difference between the two groups in survival rate. The null hypothesis of the Log-Rank test is there is no difference between the two groups. However, our test shows statistical significance with p-value of 0.0013 < 0.05. Given that our test is statistically significant, We can reject the null hypothesis of the log-rank test.

Figure 2a: Survival Probability between Sexes

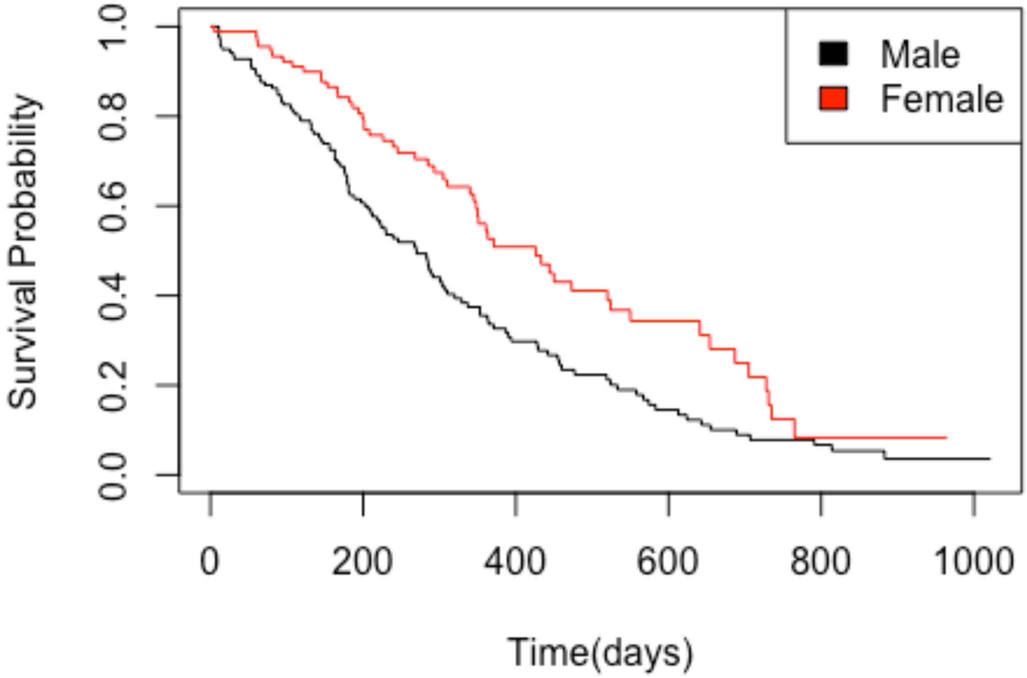
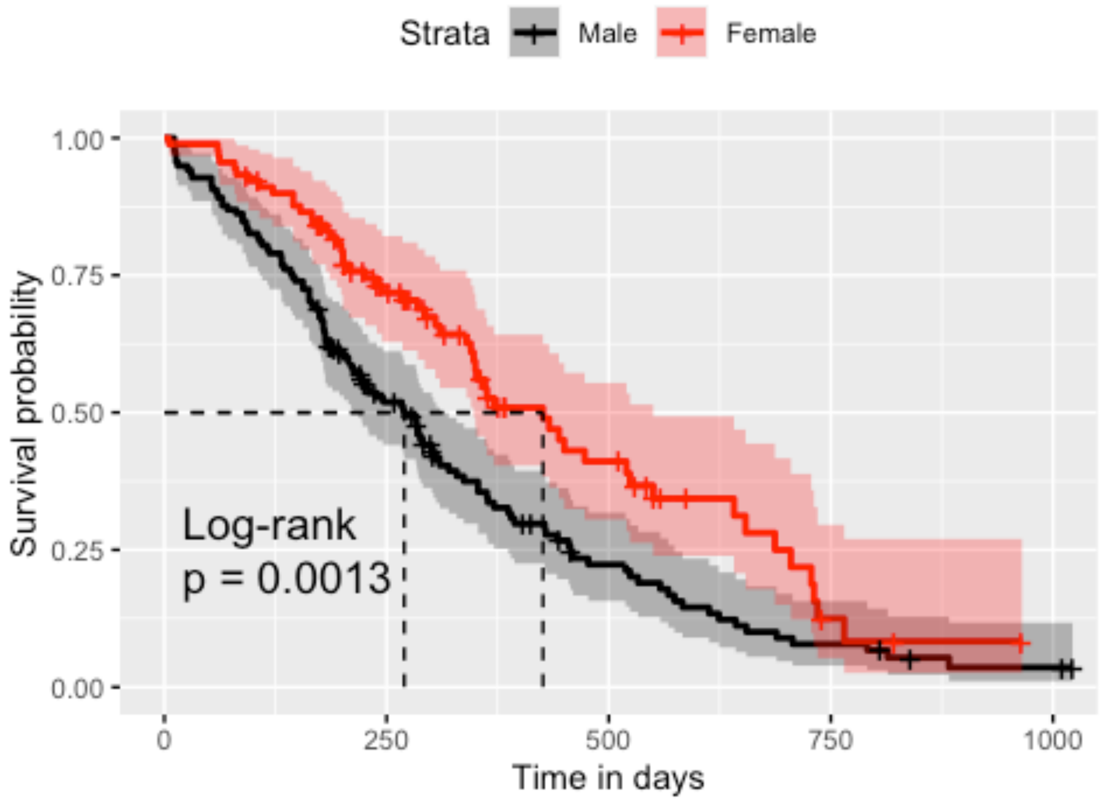


Figure 2b: Survival Probability between Sexes



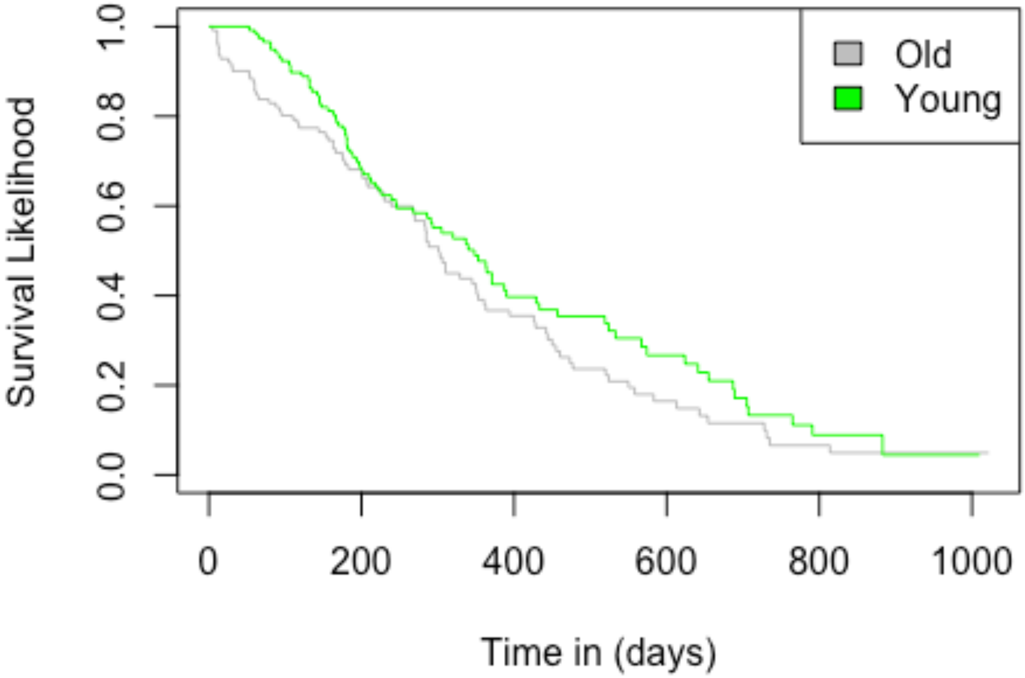
```
## Log-Rank Test Statistical Significance 0.001311165
```

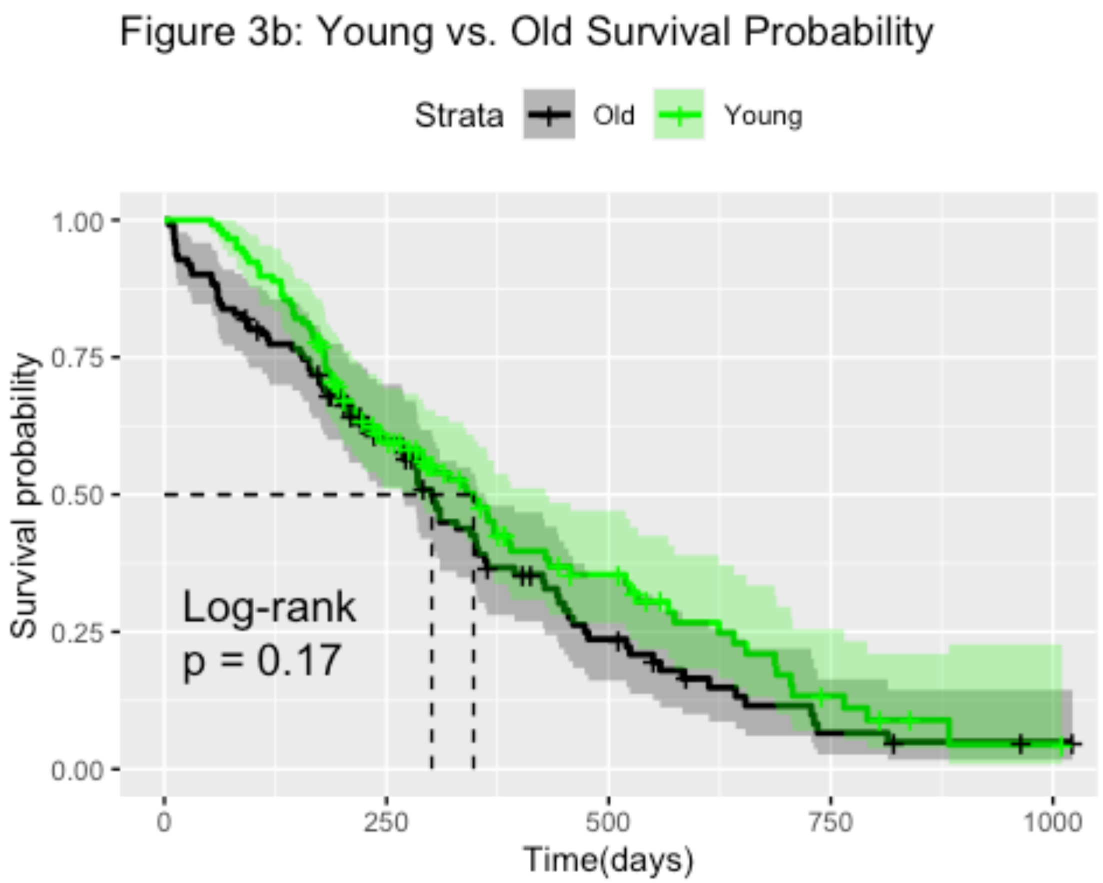
d. Is there a difference in the survival rates for the older half of the group versus the younger half? Make sure to provide a formal statistical test with a p-value and visual evidence.

From the plot, we can see that younger patients have 0.5 and less chance of survival after 80 days of Lung cancer. On the other hand, older patient shave 0.5 and less chance of survival after 85 days. Since both younger and the older patients have medians that very close to each other at 0.5 chance of survival, We suspect that there is no difference between the two groups in survival. To verify our conclusion, we used log-rank test to see if there any difference between the two groups in survival rate. The null hypothesis of the log-rank test is there is no difference between the two groups. Our test does not shows statistical significance with p-value of 0.17 > 0.05. Given that our test is not statistically significant, We can fail to reject the null hypothesis of the log-rank test and we conclude that there is no statistically significant difference between the two groups in survival over time.

```
## Call: survfit(formula = Surv(time, status) ~ young.old, data = cancer)
##
##                       n events median 0.95LCL 0.95UCL
## young.old=older   111     85    301     269     361
## young.old=younger 117     80    348     268     429
```

Figure 3a: Young vs. Old Survival Probability





```
## Log-Rank Test Statistical significance 0.001311165
```

Conclusion

The purpose of this assignment was to conduct a survival analysis on Cancer patients. we estimated the chances of a patient surviving cancer past 300 days . Also, We successfully analyzed the survival rate for different groups based on Sex and Age and determined if there is a difference between survival rate among groups using a formal statistical test with a p-value and visual evidence

###Works Cited

1.Michael, Semhar, and Christopher P. Saunders. “Survival Analysis Introduction” Chapter 11. 25 Oct. 2020, South Dakota State University, South Dakota State University. 2.Everitt, Brian, and Torsten Hothorn. A Handbook of Statistical Analyses Using n SECOND EDITION. Taylor and Francis Group LLC, 2010. 3.Jackson, Simon. “Visualising Residuals • BlogR.” BlogR on Svbtle, drsimonj.svbtle.com/visualising-residuals. 4.4.Neupane, Achal.“Survival Analysis” Achal Neupane,11 Oct.2019,achalneupane.github.io/achalneupane.github.io/post/survival_analysis/