Survival Analysis

Mohamed Ahmed

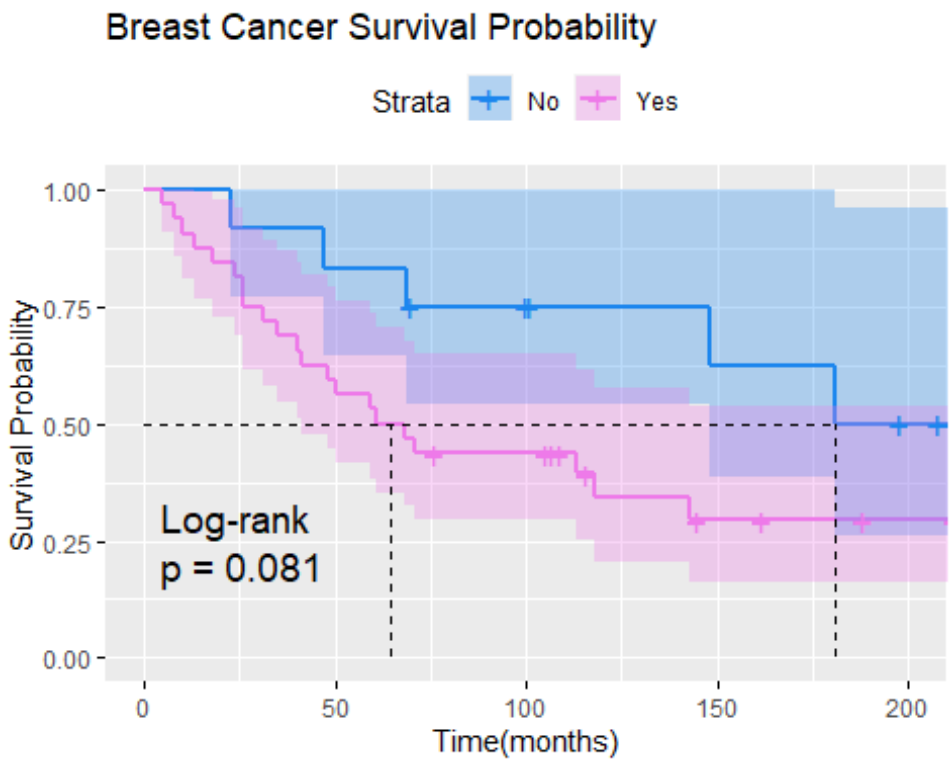
Collaboration:

Amin baabol 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.



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

## Call:
## survdiff(formula = Surv(time, event == 1) ~ metastasized, data =
mastectomy)
##
##                    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
```

c. Write one Growth rate of "variable"paragraph summarizing your findings and conclusions.

Conclusion

The Purpose of our analysis was to conduct survival analysis on **mastectomy** data. The data consists of time, events that occurred during a certain time and cancer that is classified as having metastasized or not based on a histochemical marker. Our objective was to determine if there difference between the classified cancer groups using the Kaplan-Meier curve and the log rank test. USing a censoring of event=1 will allow us to do a valid test since event=1 means that the survival time was censored.We assumed that the probability of survival is constant within each interval. Also, we assumed that only have right censoring. After fitting a model that estimates survival chances for women with metastasized cancer and for women without metastasized cancer Then we plotted the Kaplan-Meier estimate to visualize if the an apparent difference between the two groups. Finally, We conducted log-rank test to double check if there is statistically significant difference between the two groups of women with breast cancer. After inspecting the Kaplan-Meier curve for both groups, we observed the median of each group corresponding the 0.5 chances of survival. The group with metastasized cancer had 0.5 chance and less of survival after 65 months. On the other hand, the group with no metastasized cancer had 0.5 chance and less of survival after 181 months which means this group survivs longer than other group. Since the medians which correspond to 0.5 chance of survival differ greatly, We concluded that there is difference between the two groups. After conducting log rank test, We failed to reject the null hypothesis which stated that there is no difference between the two groups because the p-value of the test was insignificant (0.081>0.05).

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.

Overview

The purpose of this report is to conduct a survival analysis on *Cancer* patients. Our goal is to estimate the chances of a patient surviving cancer past *300 days*. Also, We will analyze the survival time for different groups based on *Sex* and *Age* and determine if there is a difference between survival time among groups.

Data and Model

The data that was used for this survival analysis is *Cancer* data from the *Survival* package. The data contains information such as each patient's survival time, status, age, and sex. These varibles 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% ofthe 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

For all questions of interest, we wil fit a model using the surv function to create the survival object, and analyze the rates of occurrence of events over time. Mainly, we will construct non parametric models for this analysis. Kaplan-Meier estimator is one of the non parametric technique that we will use to estimate the survival function. Also, We will use log-rank to test for differences between in survival between two groups.

Results

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

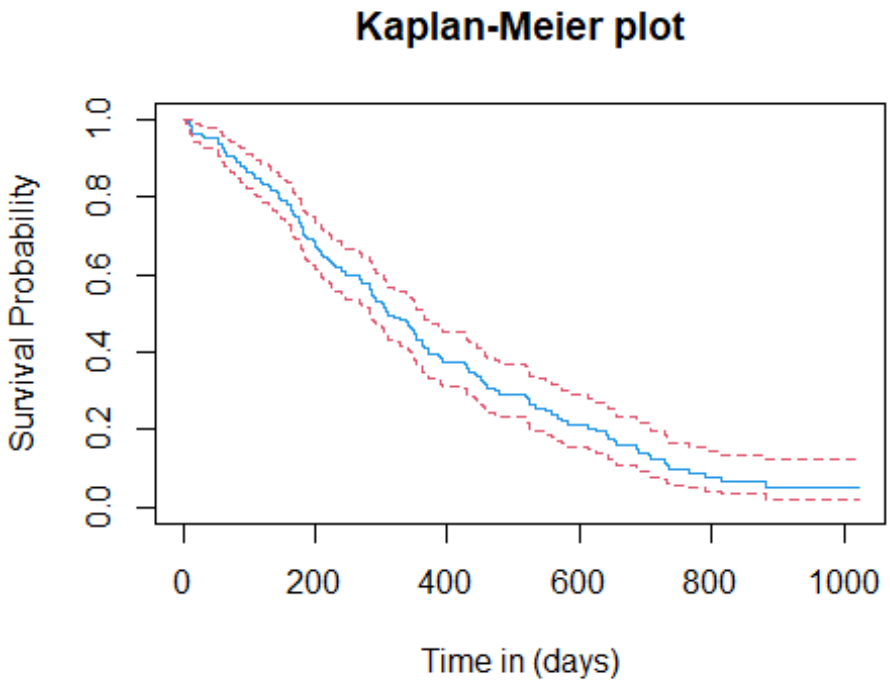
We constructed a model using surv function to estimate the chance of patient surviving past 300 days. We entered the variables time and and status == 2 (2= dead) as parameters for the model. The model estimated that a patient's chances of survival is *0.53* after 300 days.

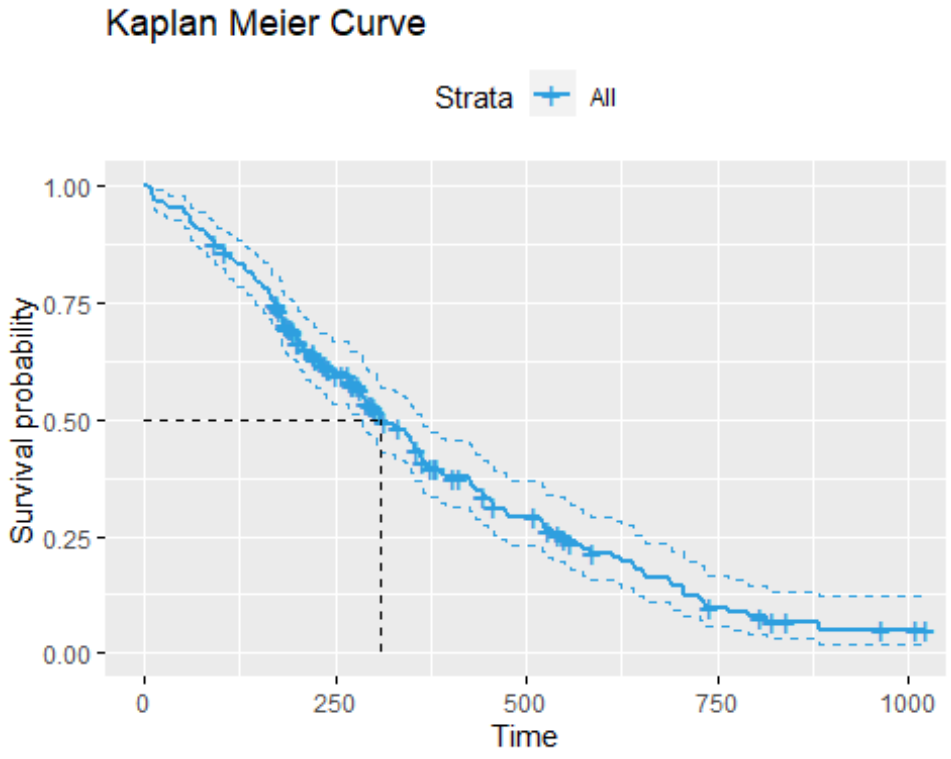
```
## Probability of survival past 300 days is 0.5247773
```

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

We constructed Kaplan-Meier curve for the entire study. The graph show that half of the patients have less tham *0.5* chances of survival after *310* days.

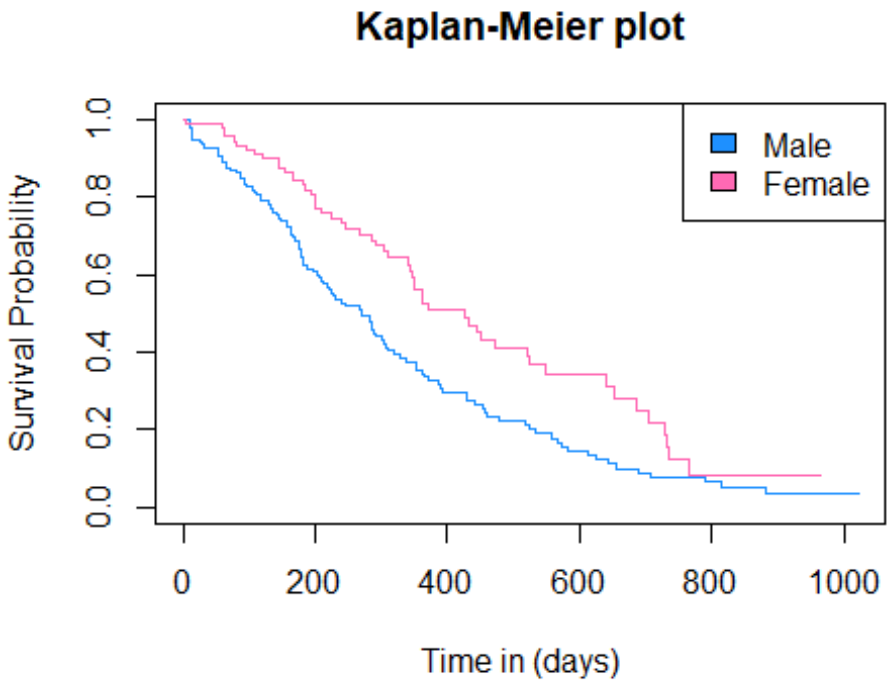
```
## Call: survfit(formula = Surv(time, status) ~ 1, data = cancer)
##
##       n  events  median 0.95LCL 0.95UCL
##     228     165     310     285     363
```

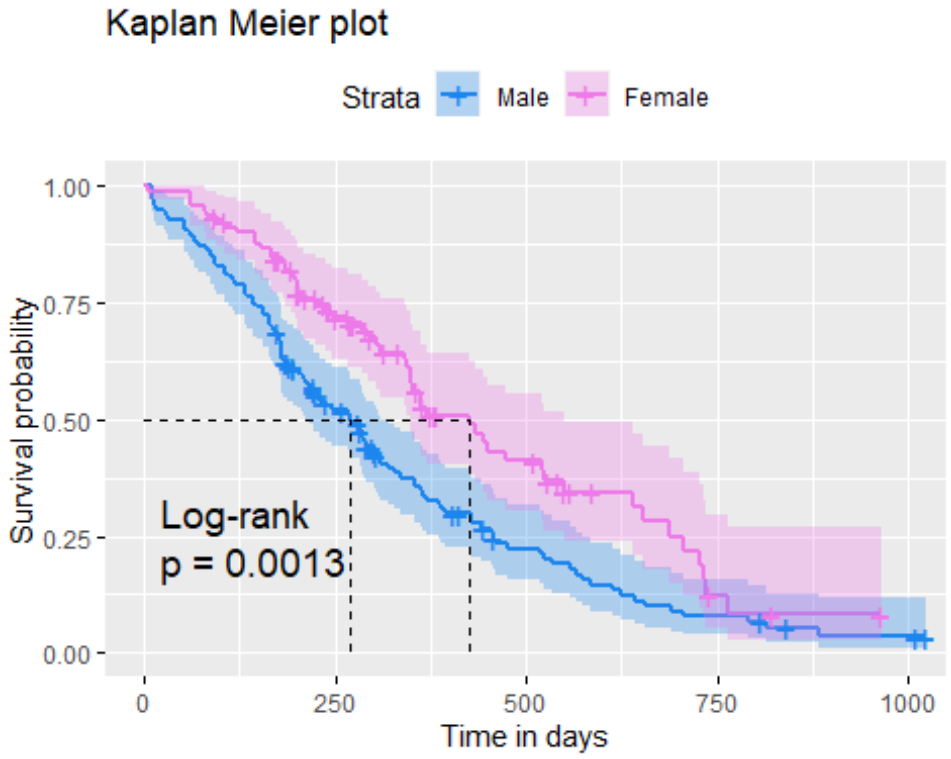




c. Is there a difference in the survival rates between males and females? Make sure to provide a formal statistical test with a p-value and visual evidence.

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 hypothsis 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.

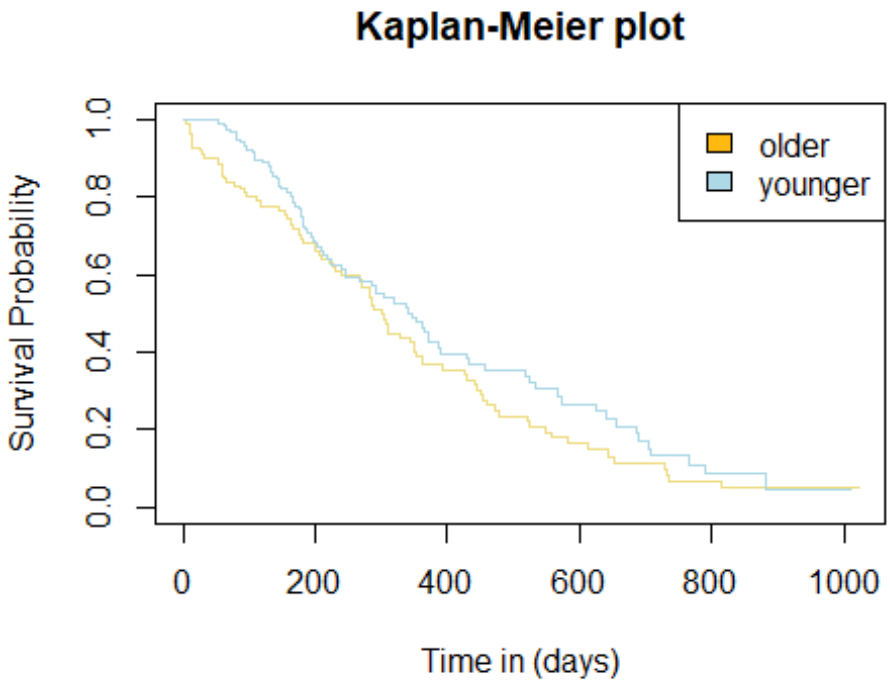


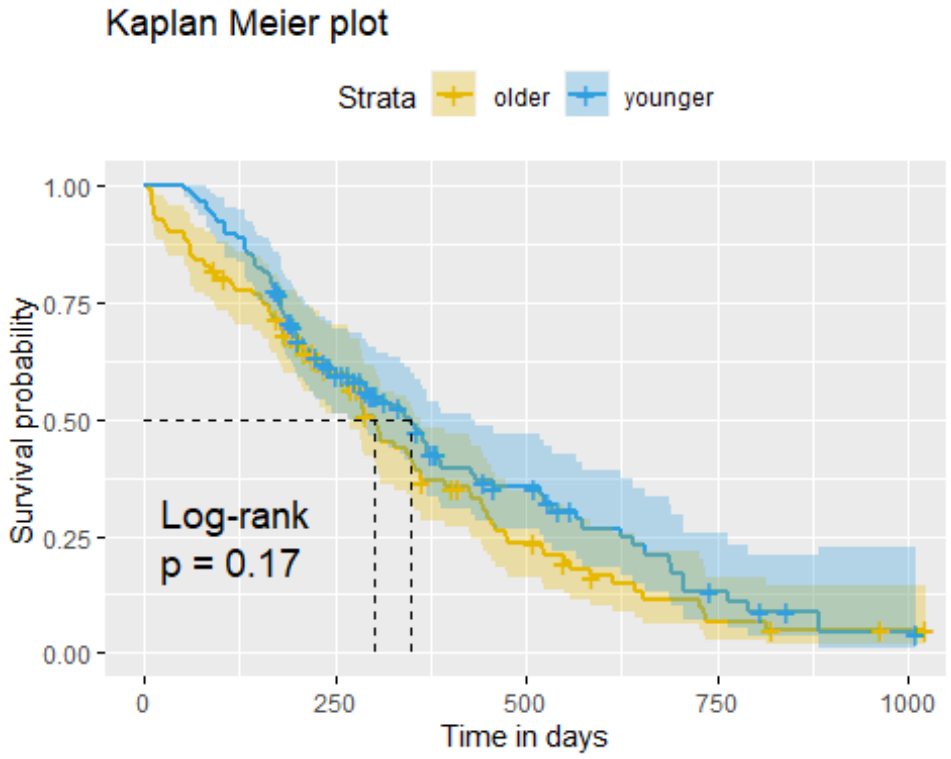



```
## Log-Rank Test Statistical Significance 0.001311165
```

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 patientsvhave 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 hypothsis 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.






```
## Log-Rank Test Statistical Significance 0.1702206
```

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 seccessfully 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

Citations

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/ Survival Analysis Basics. (n.d.). Retrieved October 28, 2020, from http://www.sthda.com/english/wiki/survival-analysis-basics