Homework #7

Justin Robinette

October 9, 2018

No collaborators for any problem

Problem #1, Part A: An investigator collected data on survival of patients with lung cancer at Mayo Clinic. The investigator would like you, the statistician, to answer the following questions and provide some graphs. Use the **cancer** data located in the **survival** package.

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

Results: *Figure 1.1* shows the probability that someone will survive beyond 300 days as **0.5306081**. As interest to me, I also included *Figure 1.2* showing the proportion of subjects in the study that survived beyond 300 days. This proportion is **0.3991228**. The difference is of interest to me using this function.

Figure 1.1: Probability of Surviving Past 300 Days

Probability
0.5306081

Figure 1.2: Proportion of Subjects Surviving Greater than 300 Days

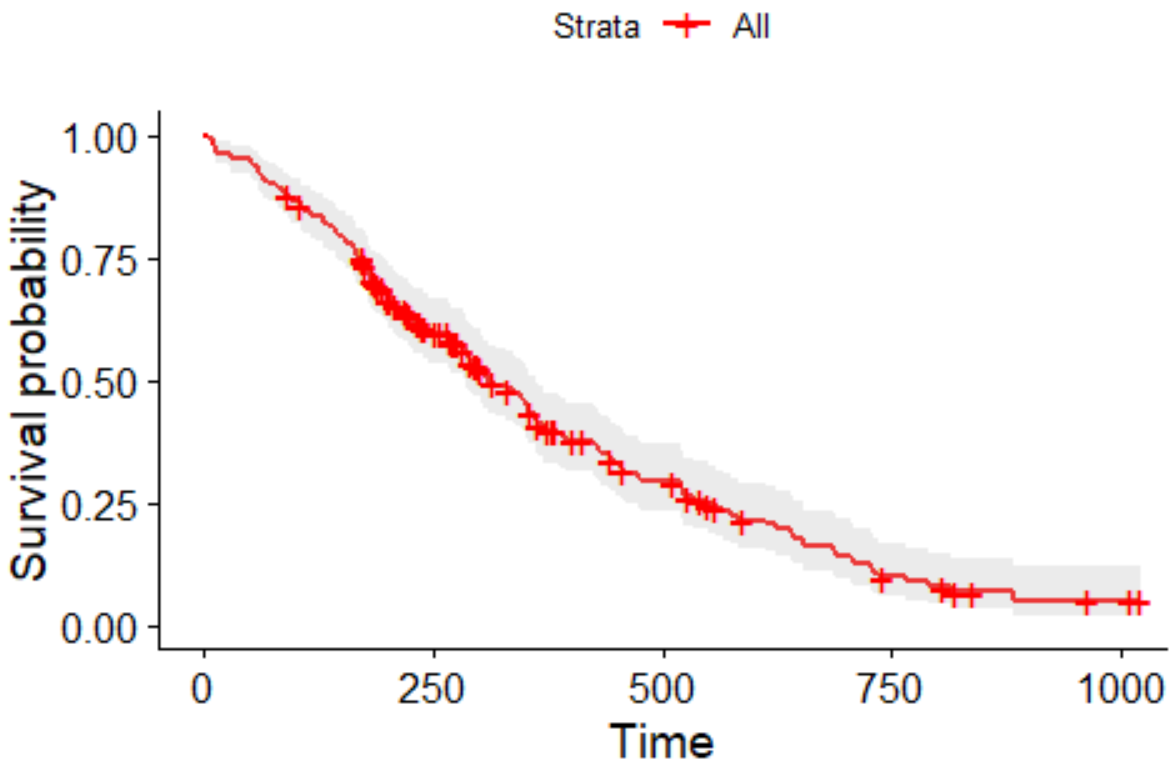
Proportion
0.3991228

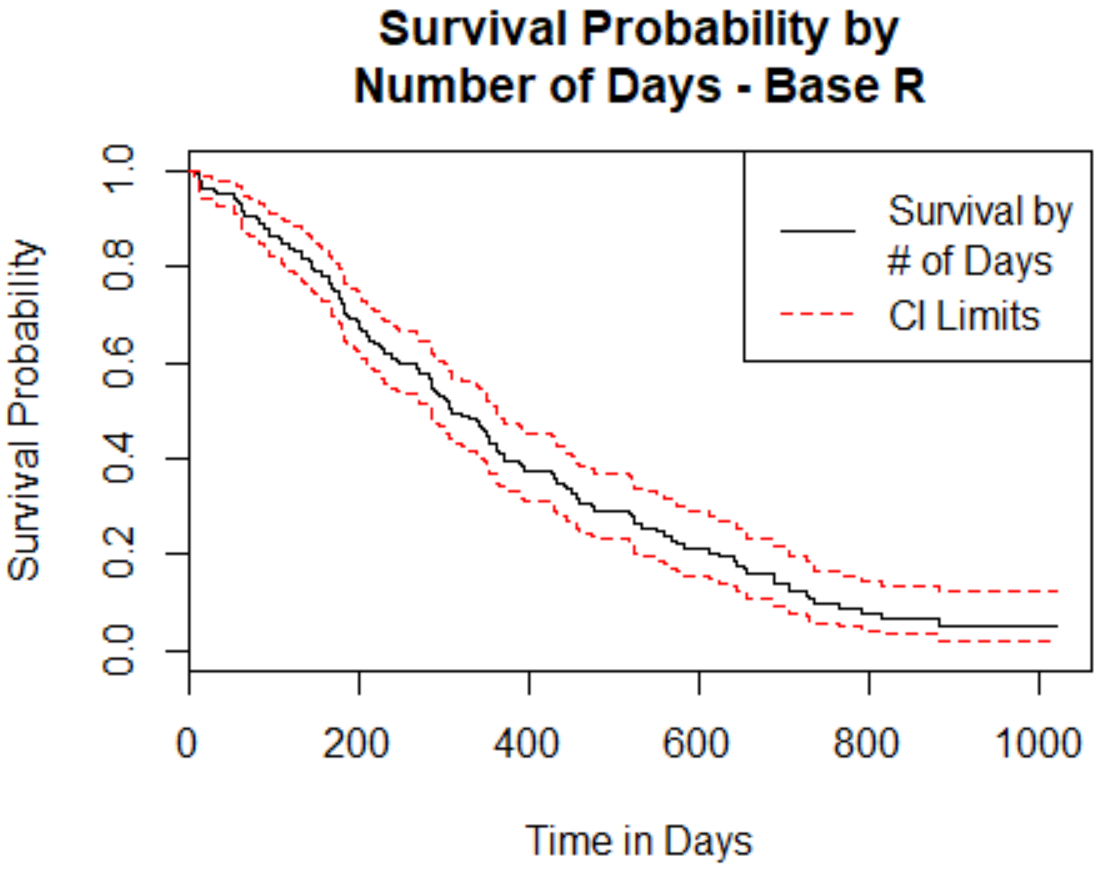
Problem #1, Part B: Provide a graph, including 95% confidence limits, of the Kaplan-Meier estimate of the entire study.

Results: *Figure 1.3* shows a plot of the probability of surviving by the number of days using the **Kaplan-Meier** estimate of the entire study.

The base R plot shows the estimate a little better, I feel. The black line shows the actual probability of the subjects in the dataset. The red lines show the **95% confidence interval range** (both upper 2.5% and lower 2.5%).

Figure 1.3: Survival Probability
by Number of Days





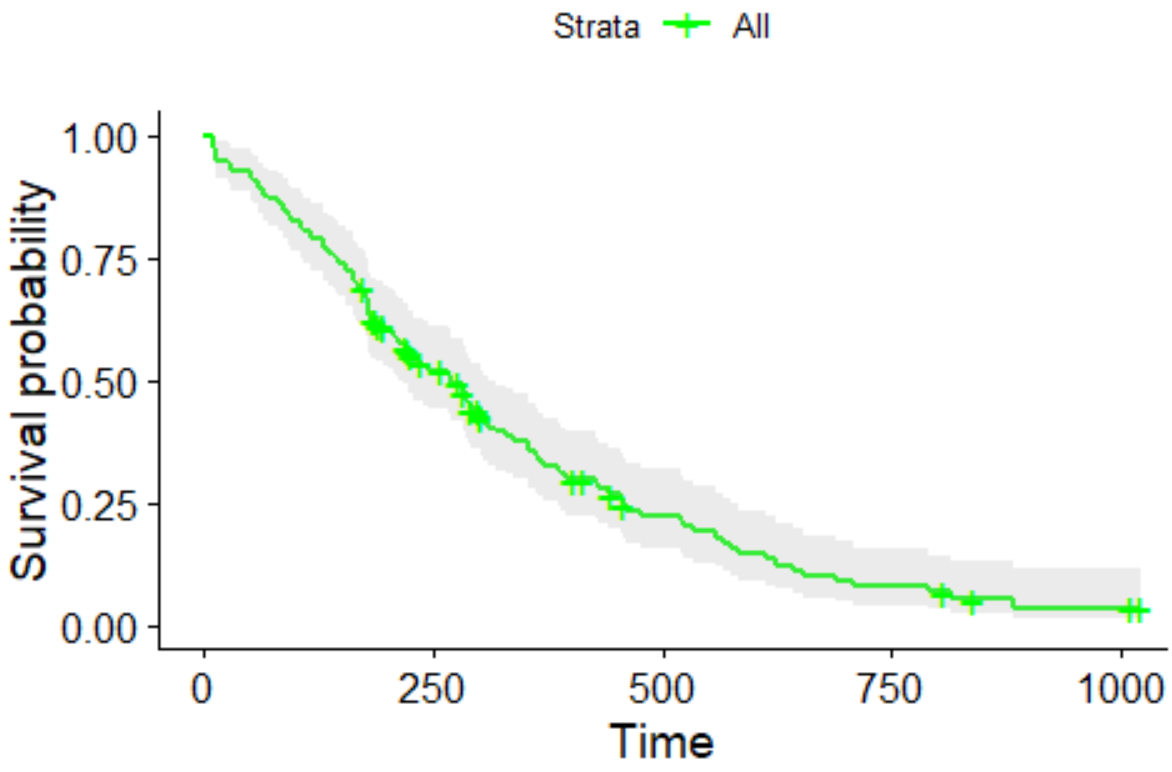
Problem #1, Part C: Is there a difference in the survival rates between males and females? Provide formal statistical test with p-value and visual evidence.

Results: *Figure 1.4* and *Figure 1.5* show the plots, by gender, of the probability of survival by number of days. *Figure 1.5* shows a definite increase in probability of survival for women over men. Base R plots are included, per assignment instructions.

Figure 1.6 is a follow-up to *Figure 1.1* above showing the probability of surviving 300 days by men and women. *Figure 1.1* had shown the probability of the population is **0.5306081**. *Figure 1.6* summarizes the probability divided for men and women. As we see, women have a much higher probability of surviving greater than 300 days than men.

Figure 1.7 shows the P-Value, as requested in this exercise, determining the statistical significance of the difference in men and women and their chances of survival. At a p-value of **0.001046**, it is safe to say that the difference in survival for men and women is statisticall significant.

Figure 1.4: Male Survival Probability by Number of Days



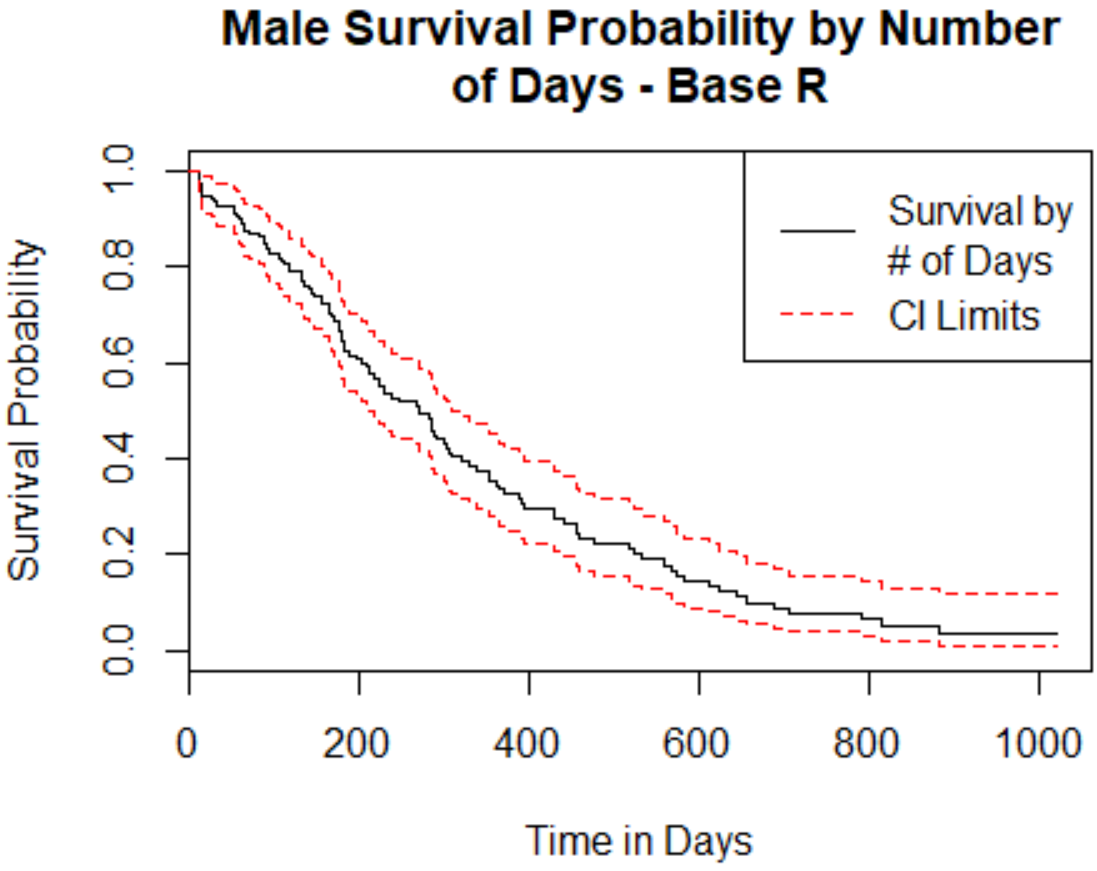
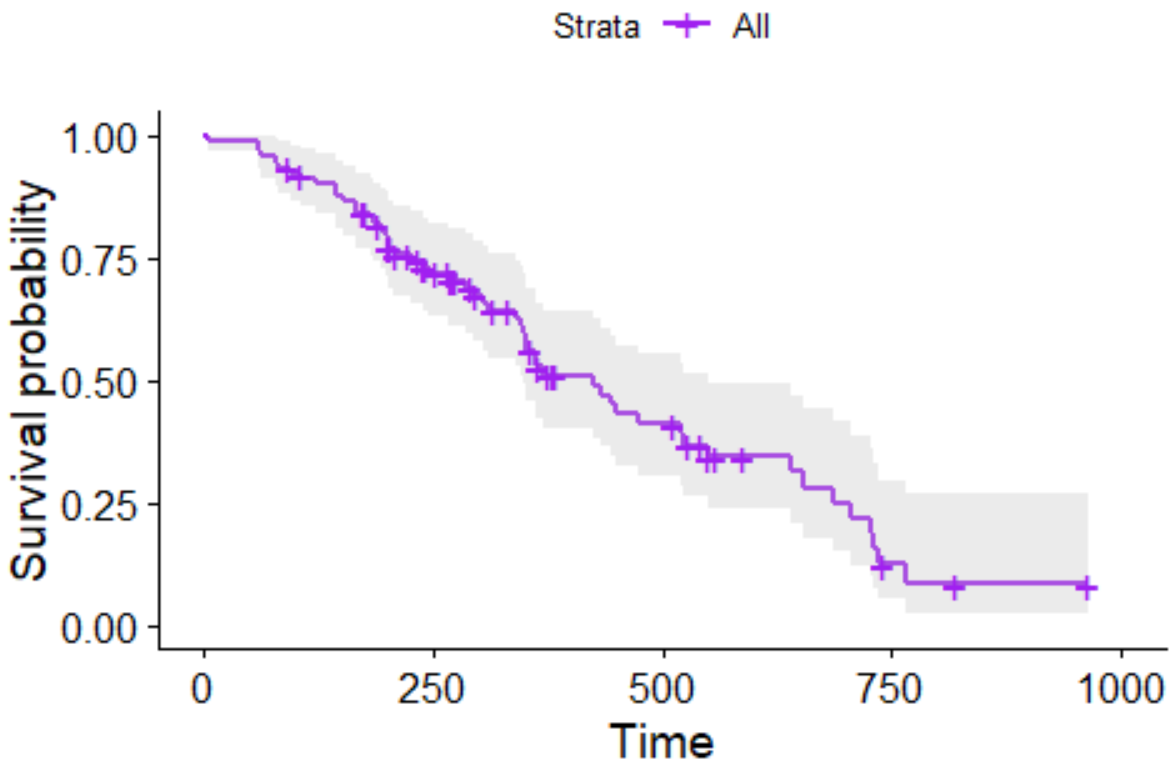


Figure 1.5: Female Survival Probability by Number of Days



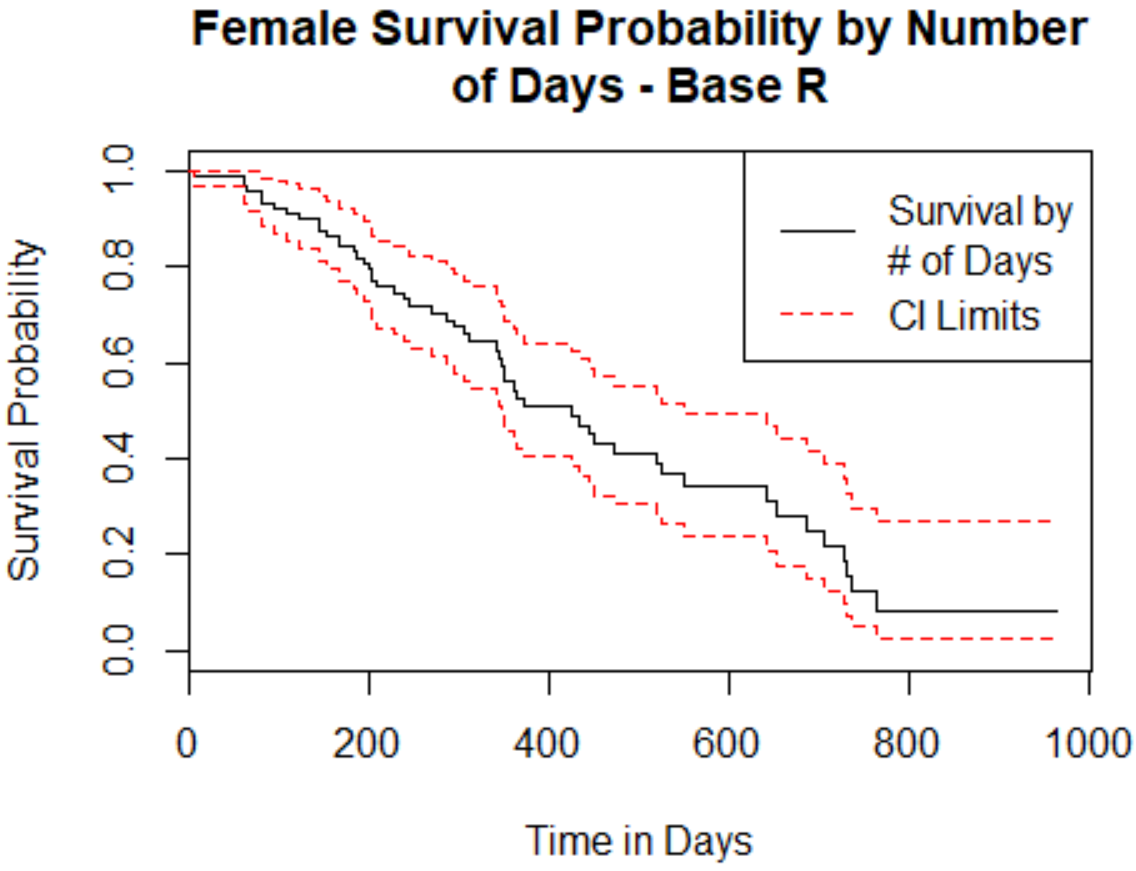


Figure 1.6: Probability of Surviving Past 300 Days by Gender

Male Probability	Female Probability
0.4411	0.6742

Figure 1.7: P-Value of Survival Difference between Men and Women

P-Value
0.001046

Problem #1, Part D: Is there a difference in the survival rates for the older half of the group versus the younger half?

Provide a formal statistical test with p-value and visual evidence.

Results: The question asks us to split the dataset into older and younger half. Since there are 11 people with the median age of 63, it was not going to be possible to split the dataset evenly into older and younger without making a somewhat arbitrary decision. For example, removing all people with the median age would not leave us with two evenly dispersed groups of "older" and "younger" participants.

For this reason, I made the decision to set anyone over the median age as "older" and anyone at the median age or below as "younger". This gets us close to an even split of 111 "older" subjects and 117 "younger" subjects. I've summarized this breakdown in *Figure 1.8* but I felt it was an important clarification before moving forward with the exercise.

Figure 1.9 and *Figure 1.10* visually depict the survival probabilities for older and younger participants by number of days of survival. We see more similarity between older and younger observations than we saw for men and women in *Figure 1.4* and *Figure 1.5*.

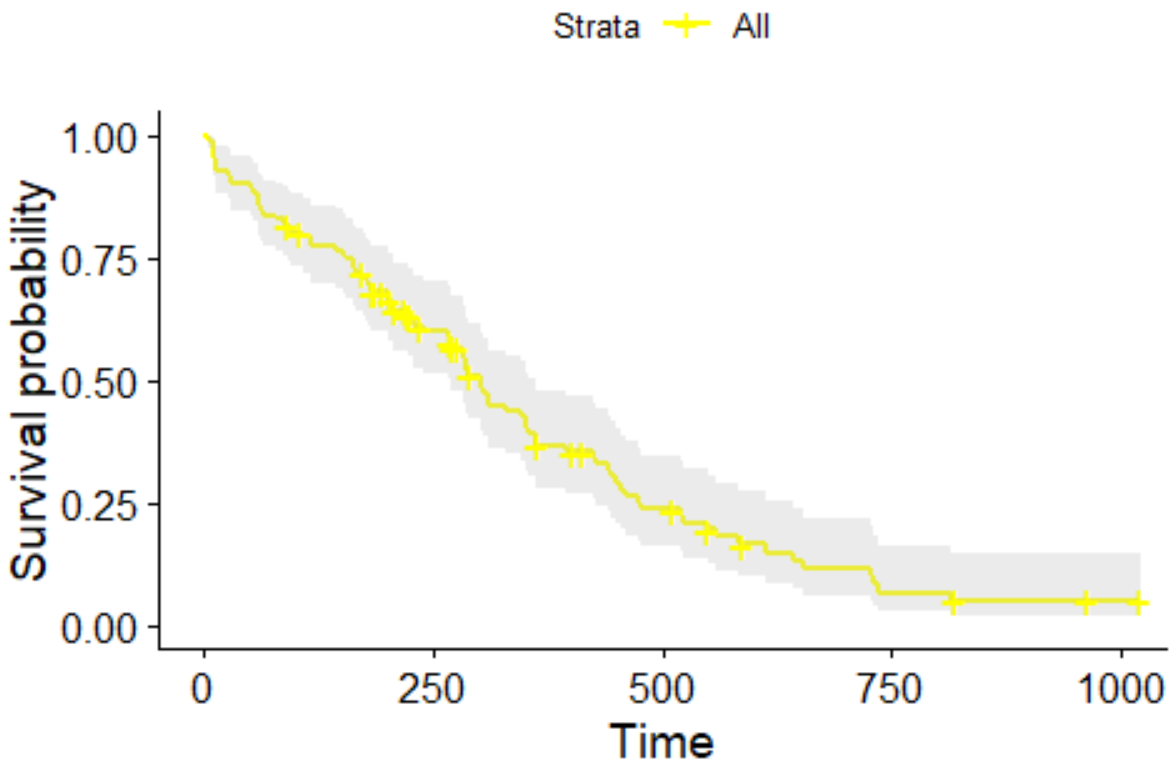
Figure 1.11 shows the likelihood of survival beyond 300 days for participants above the median age and at or below the median age of the study. We see a slight advantage for younger participants, but not as pronounced as what we saw in comparing men and women.

Figure 1.12 shows the P-Value of the statistical significance between young and old participants and their likelihood of survival. With a p-value of **0.168915** we can say that we do not have evidence to reject our Null Hypothesis at an alpha of 0.05 or 0.1. This means we cannot reject the hypothesis that states that there is no significant difference in survival likelihood for older and younger participants.

Figure 1.8: Summary of Older/Younger Breakdown

Older Obs	Younger Obs	Obs at Median
111	117	11

Figure 1.9: Older Survival Probability by Number of Days



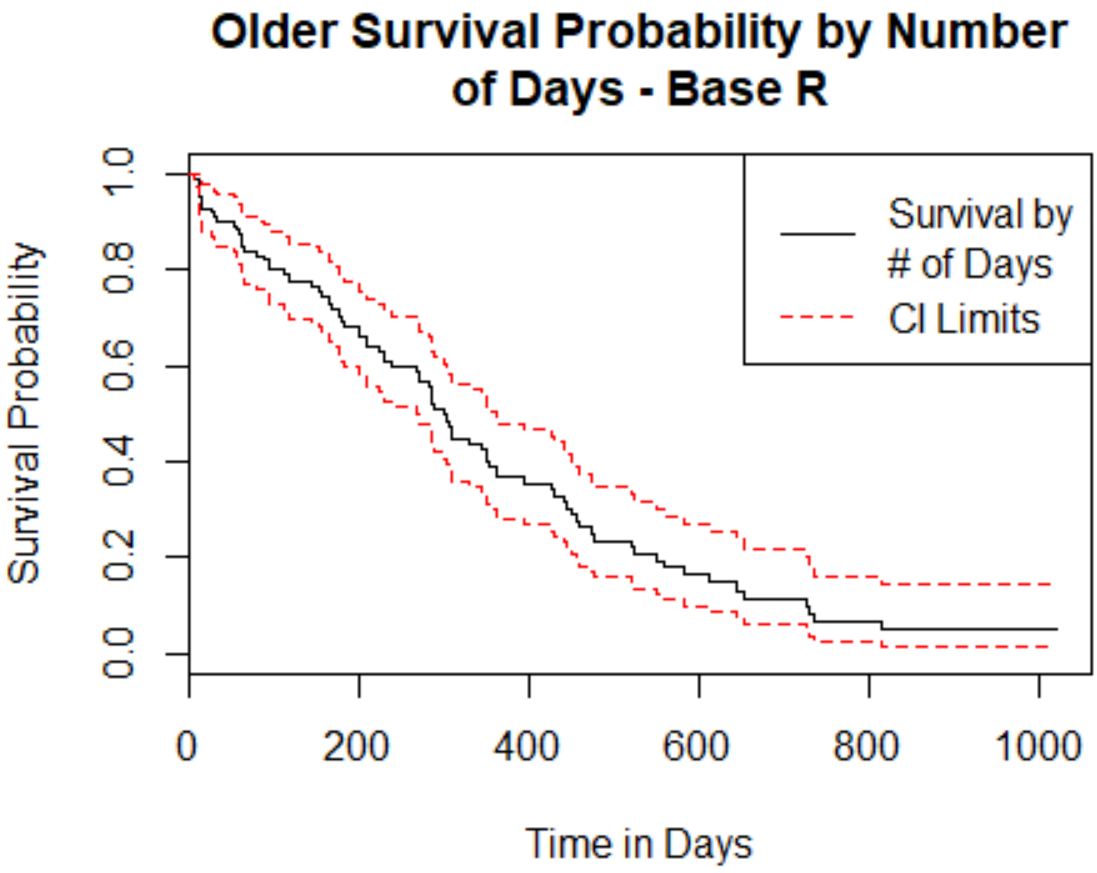
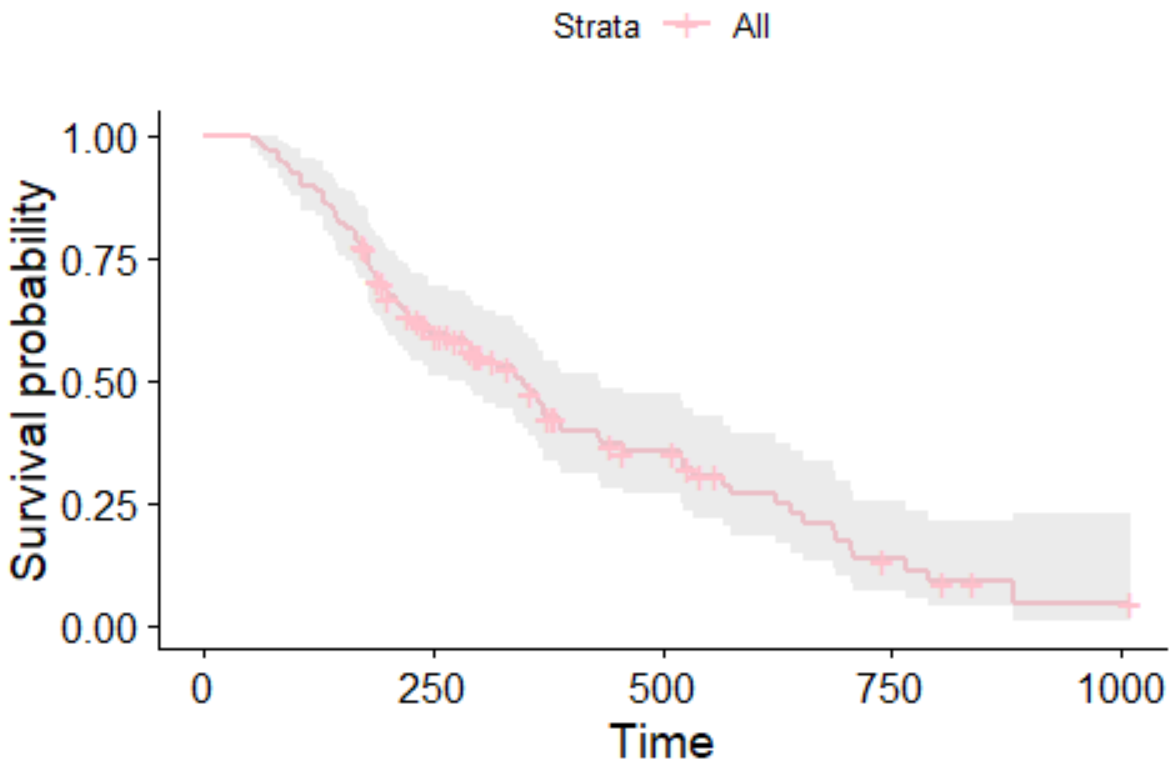


Figure 1.10: Younger Survival Probability by Number of Days



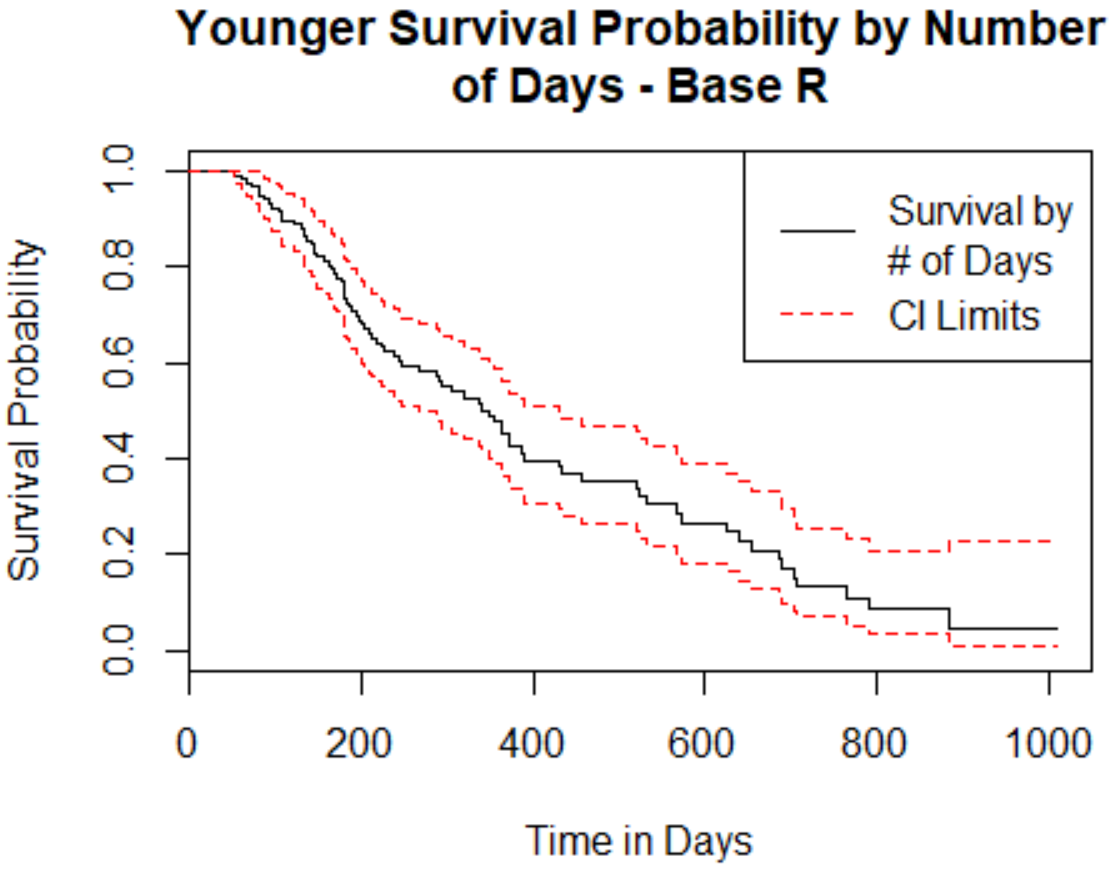


Figure 1.11: Probability of Surviving Past 300 Days by Age

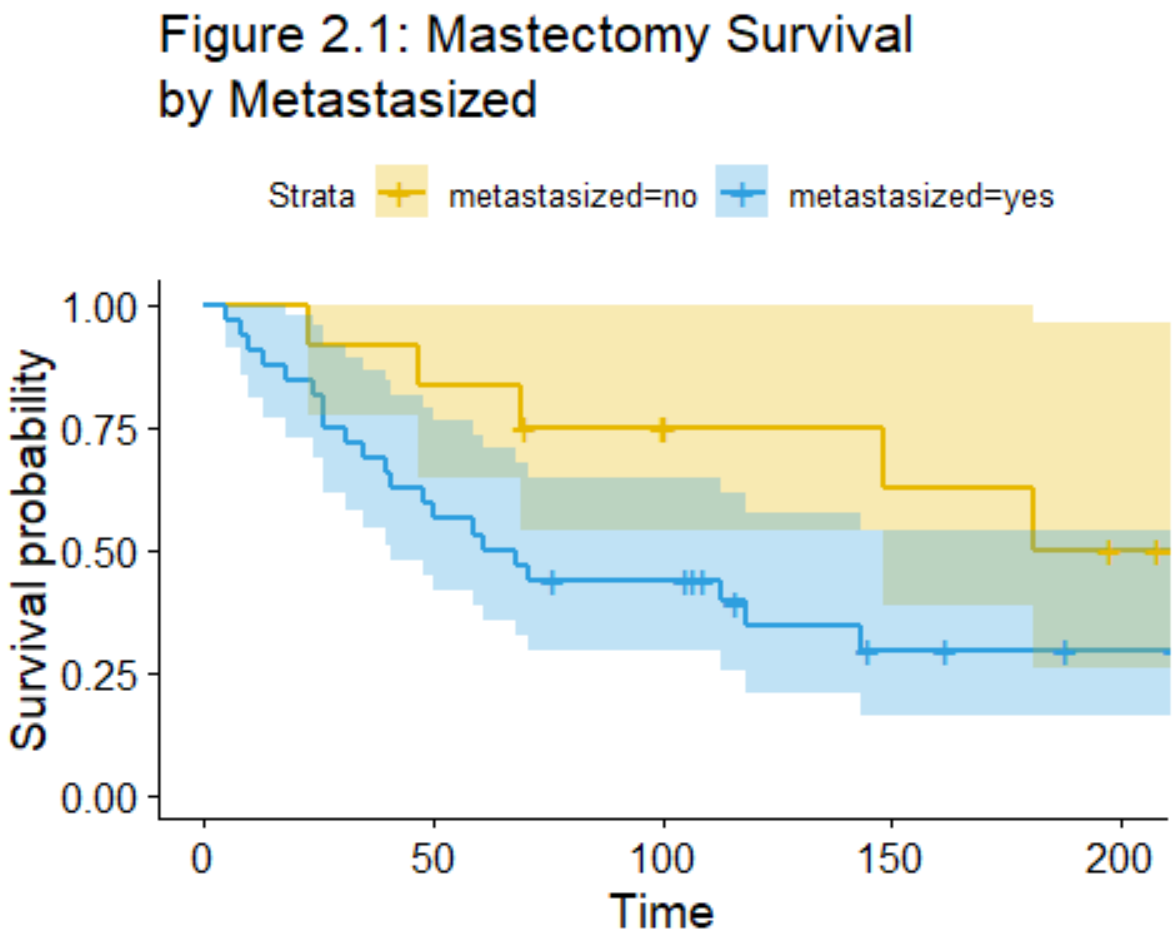
Older Probability	Younger Probability
0.5089	0.5513

Figure 1.12: P-Value of Survival Difference by Age

P-Value
0.168915

Problem #2, Part A: A healthcare group has asked you to analyse 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 classiffied 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. Do the following:

Plot the survivor functions of each group only using GGPlot, estimated using the Kaplan-Meier estimate.



Problem #2, Part B: Use a log-rank test to compare the survival experience of each group more formally. Only present a formal table of your results.

```
        Asymptotic Two-Sample Logrank Test
data:  Surv(time, event == 2) by metastasized (no, yes)
                Z = 1.8667, p-value = 0.06194
  alternative hypothesis: true theta is not equal to 1
```

Problem #2, Part C: Write one paragraph summarizing your findings and conclusions.

Results: From *Figure 2.1* we see that the survival probability is higher if the cancer did not metastasize. The range, however, for the 95% confidence interval is also larger when the cancer did not metastasize. The log-rank test results show a higher p-value than I would've expected. Assuming an alpha of 0.05, we would not have evidence to reject the Null Hypothesis. The Null Hypothesis would state that there is not a significant difference in survival probabilities depending on whether a patient had the cancer metastasize. If we were using a larger alpha, say 0.1, then we would have evidence to reject the Null since the p-value is **0.6194**