Homework 7

Alex Soupir

October 30, 2019

Packages: HSAUR3, ISwR, survival, coin, party

Collaborators:

Please do the following problems from the text book R Handbook and stated.

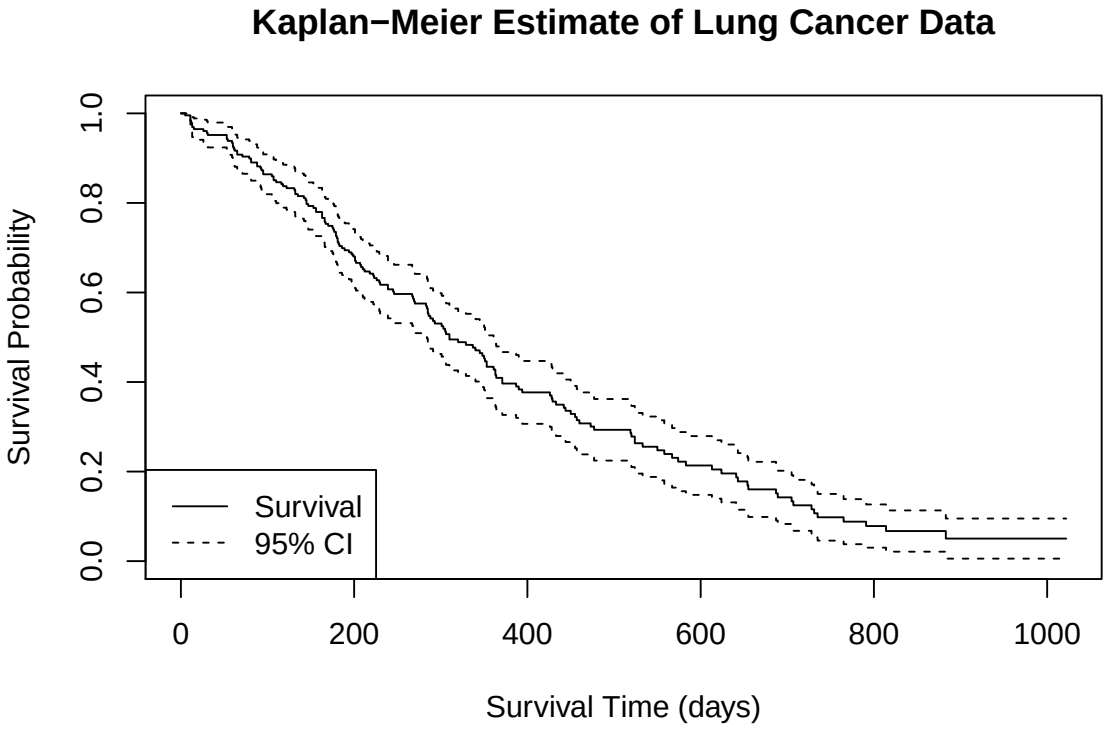
1. An investigator collected data on survival of patients with lung cancer at Mayo Clinic. The investigator would like you, the statistician, to answer the following questions and provide some graphs. Use the **cancer** data located in the **survival** package.

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

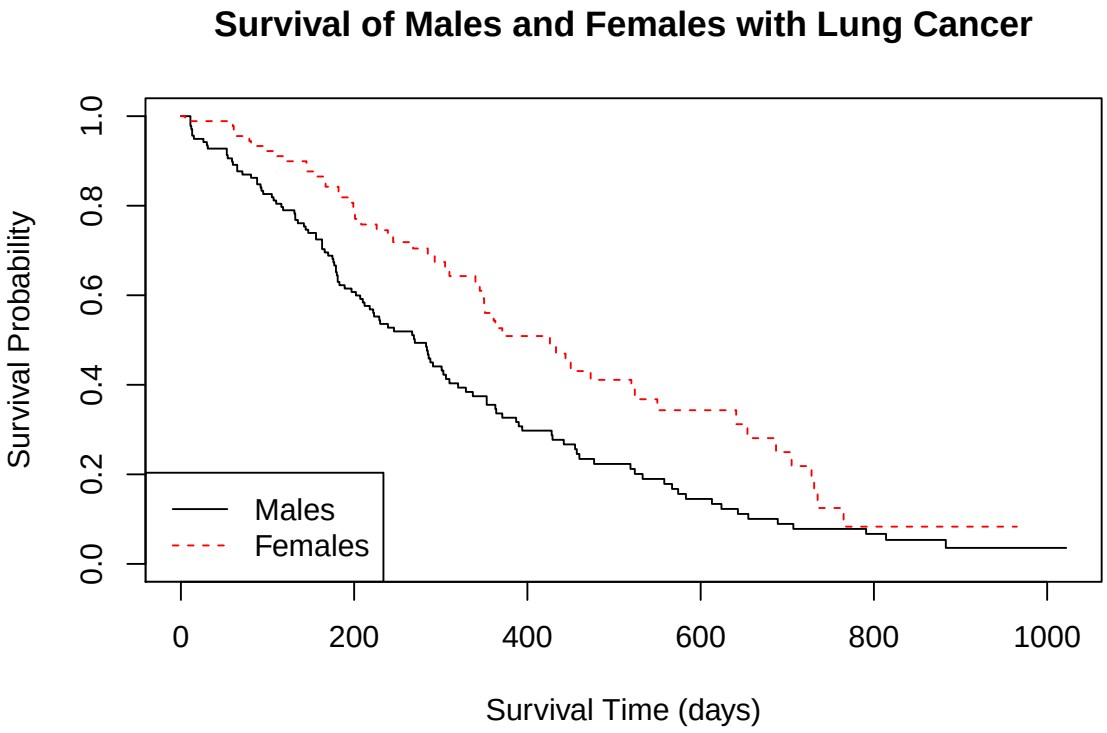
```
## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = cancer,
##     conf.int = 0.95, conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     92     101    0.531  0.0346        0.463        0.599
```

   Dead patients were given a status of 2, so the status in *Surv* has to be == 2. The probability thta someone will live past 300 days is 0.531.

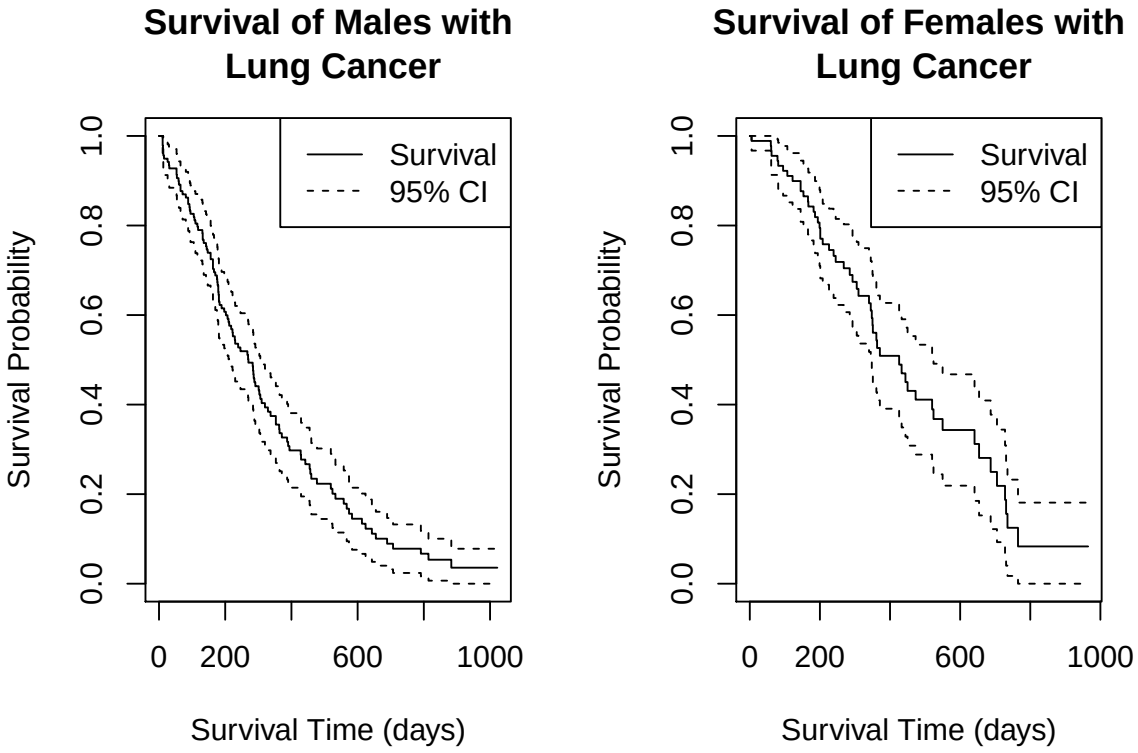
   b. Provide a graph, including 95% confidence limits, of the Kaplan-Meier estimate of the entire study.



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




```
##
## Probability of a male living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = m1c, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     49      74   0.4411  0.0439       0.3550        0.527
##   750      7      35   0.0781  0.0276       0.0239        0.132

##
## Probability of a female living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = f1c, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     43      27    0.674  0.0523       0.5717        0.777
##   750      3      25    0.125  0.0549       0.0173        0.232
```

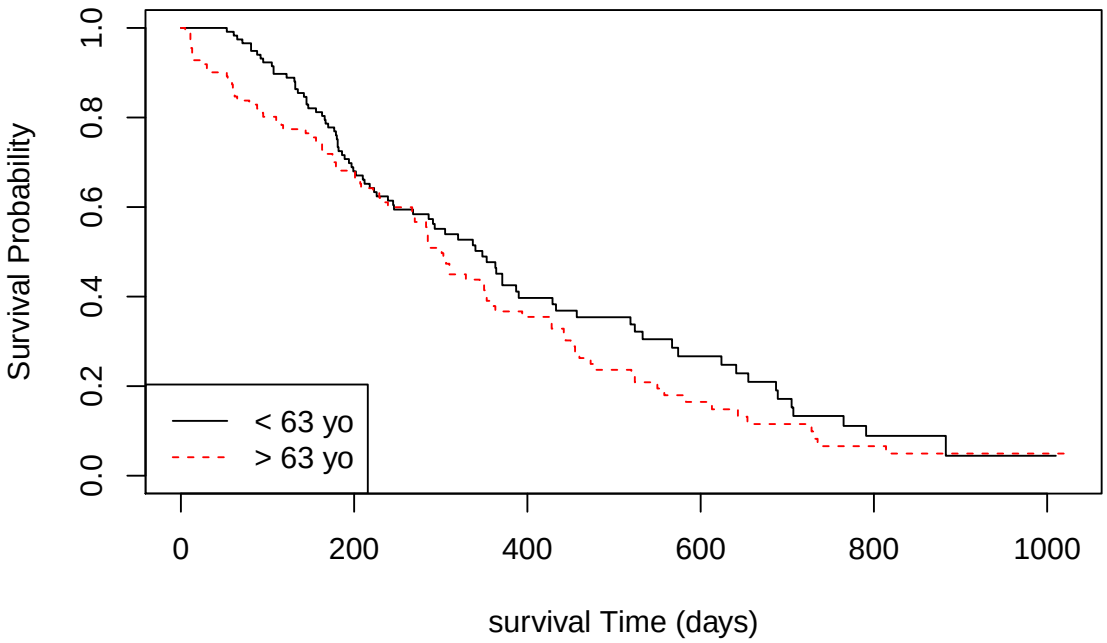



```
##
## Chisq test between males and females:

## 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
```

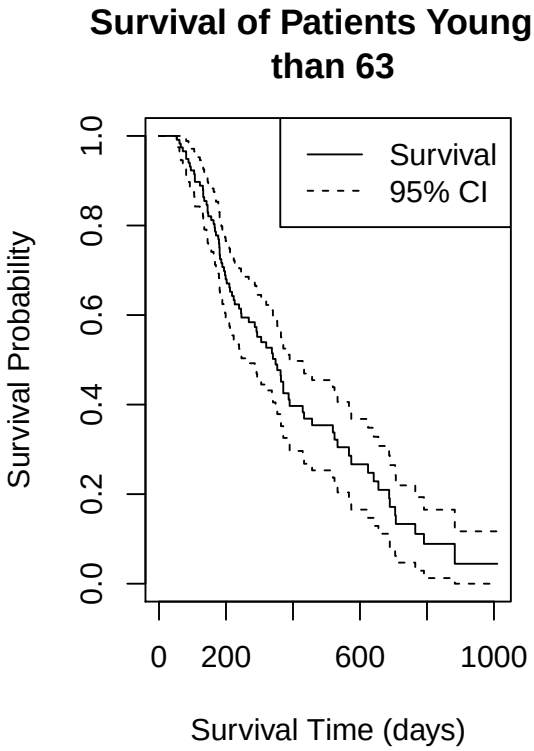
First a survival plot was created for males and females; females appear to have a slightly higher probability of survival until around 750 days. Next, the probability of surviving past 300 days and 750 days. Males probability of surviving past 300 days is 0.4411 whereas females surviving past 300 days is 0.674. Surviving past 750 days was closer between males and females with a probability of 0.0781 and 0.125, respectively. These values don't have any statistical significance associated using *summary()* so a statistical test is needed which can be done with *survdiff()*. This computes a chisq of males vs females survival rates, which resulted in a p-value of 0.001, indicating that there is a difference in survival between the genders.

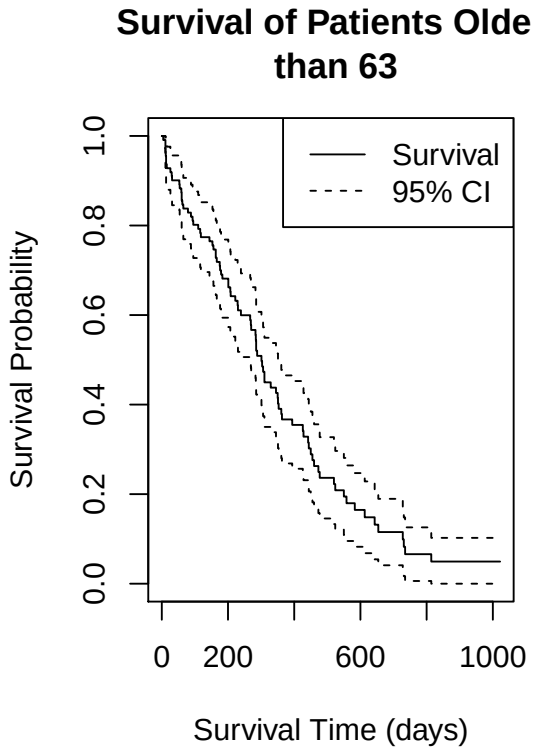
d. Is there a difference in the survival rates for the older half of the group versus the younger half? Provide a formal statistical test with a p-value and visual evidence.

```
## Finding median age of cancer patients:

## [1] 63
```




```
##
## Probability of younger than 63 living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = l1d, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     49      50    0.551  0.0478       0.4576        0.645
##   750      6      27    0.133  0.0440       0.0471        0.220

##
## Probability of those older than 63 living past 300 and 750 days:

## Call: survfit(formula = Surv(time, status == 2) ~ 1, data = g1d, conf.int = 0.95,
##     conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   300     43      51   0.5089  0.0501      0.41060        0.607
##   750      4      33   0.0659  0.0306      0.00601        0.126
```






```
##
## Chisq test between males and females:

## Call:
## survdiff(formula = Surv(time, status == 2) ~ agecat, data = cancer)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## agecat=<63 117       80     88.8     0.865      1.88
## agecat=>63 111       85     76.2     1.007      1.88
##
##  Chisq= 1.9  on 1 degrees of freedom, p= 0.2
```

Similarly to *Part C* a plot was created first to compare the 2 groups. A new column was created to distinguish between the oldest half of patients and the youngest half of patients. The median age was found to be 63 and that value was used as the split age for oldest and youngest. The first plot showing the 2 different groups shows that the 2 groups don't deviate too much from each other. The survival rates for the 2 groups was found using *survfit()* for times of 300 and 750 days. Patients younger than 63 years old have a probability of surviving past 300 days of 0.551 and patients older than 63 have a probability of 0.509 of surviving past 300 days. For 750 days, the survival probability of patients younger and older than 63 years of age is 0.133 and 0.066, respectively. To determine significance between the 2 age groups, *survdiff()* was used to compute a chi-squared p-value of 0.2, which means that there isn't a significant difference in survival rates between the older half and the younger half of lung cancer patients.

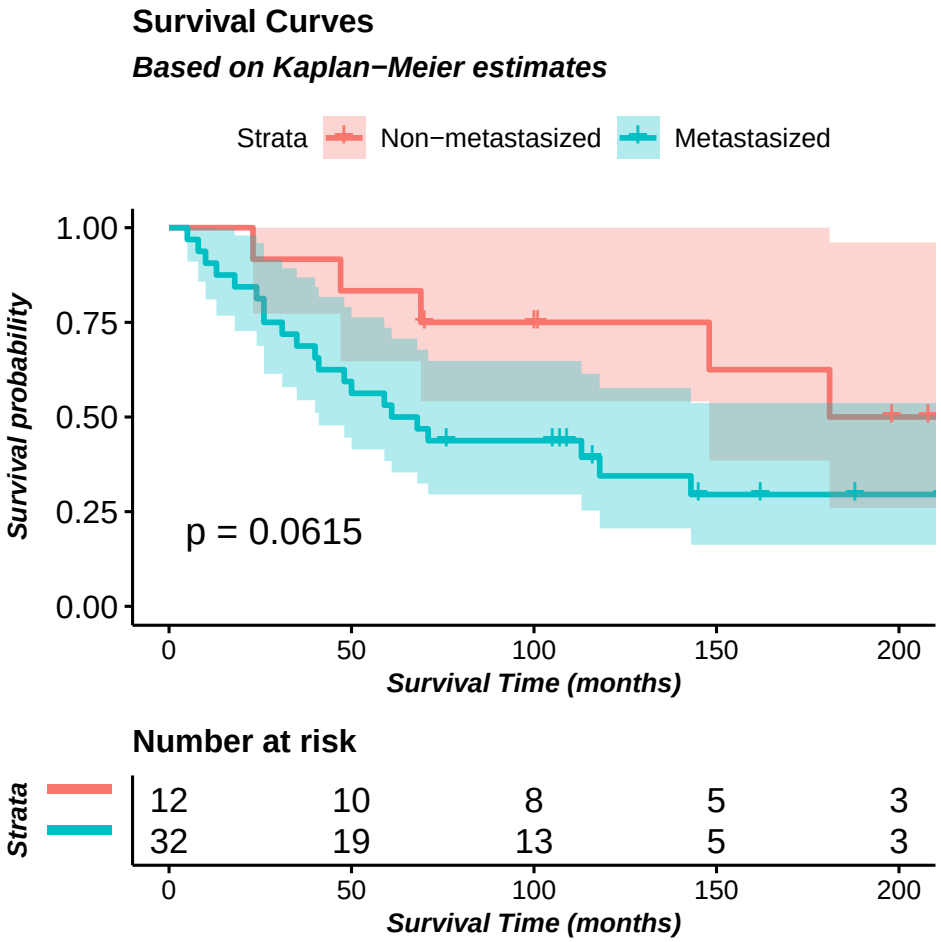
2. A healthcare group has asked you to analyse the **mastectomy** data from the **HSAUR3** package, which is the survival times (in months) after a mastectomy of women with breast cancer. The cancers are classified as having metastasized or not based on a histochemical marker. The healthcare group requests that your report should not be longer than one page, and must only consist of one plot, one

table, and one paragraph. Do the following:

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

b. Use a log-rank 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.




```
## Log-Rank Test Results:

##                                      Formula Z.value P.value
## 1 Surv(time, event == TRUE) by metastasized  1.8667 0.06146
```

A plot was created using a GGPlot wrapper (*survminer*) for the survival of patients with and without metastasized breast cancer and plotted with the 95% confidence interval. A risk table was also given from GGPlot showing the number at risk for each given time point (x50 weeks). The non-metastatic breast cancer patients appear to have a higher survival probability throughout the time course, however, the confidence intervals of the 2 groups overlap. In order to draw a definitive conslusion on whether patients with non-metastatic cancer have similar survival rates as patients with metastatic cancer, a Log-Rank test was conducted. The Log-Rank test compares the entire survival curve between the groups and tells us whether the curves are identical or not. The p-value of the Log-Rank test is 0.0615, letting us know that the curve for metastatic and non-metastatic patiets do not have statistically different survival rates across 225 weeks.

Resources Used:

- rdocumentation.org
- stackexchange.com
- rpkgs.datanovia.com/survminer
- sphweb.bumc.bu.edu (Boston University)