ST525 HW 3

Nora Quick

Question 1

a) I would say this case is left censored because we do not know when they initially became infected and tested positive. All we know is that at the initial test they were positive.

b) Since we are looking for a positive test this is equal to "being alive at the end of a study" as the person outlasted the tests for a positive I would say this is right censored.

c) This case is interval censored due to the fact that the patient became positive within the interval of the second test and the third test but we do not know exactly when between the two intervals.

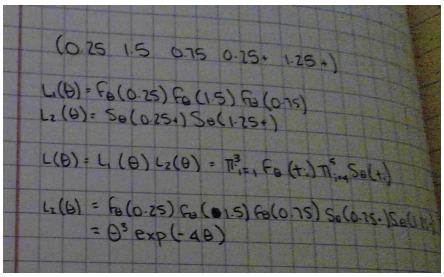
Question 2

a) This is a case of right censoring because the patient "outlasted" the study. They were healthy at the beginning and did not have breast cancer at the end of the study.

b) This is a case of interval censoring because the patient developed breast cancer sometime between the 12th and 15th year but we do not know exactly when. This is a case of right censoring because the person died of an unrelated cause before the study was over.

c) I believe this could be classified as a case of right censoring because they moved away so they could not finish the study but they never developed breast cancer while in the study.

Question 3



Question 4

```
# Part a
pSmoking <- read.csv('pharmacoSmoking-new.csv')
head(pSmoking)
```

```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26             1
## 2 113  14       1   2  41      1    4          2           27             1
## 3  39   5       1   1  25      0    4          2           12             1
## 4  80  16       1   1  54      1    4          1           39             1
## 5  87   0       1   1  45      1    4          2           30             1
## 6  29 157       0   1  43      1    2          1           30             1
##    admitdate      fdate priorAttempts longestNoSmoke
## 1 11/20/2005 12/31/2005             0              0
## 2  6/16/2005  6/30/2005             3             90
## 3   5/9/2005  5/14/2005             3             21
## 4 10/26/2005 11/11/2005             0              0
## 5  9/27/2005  9/27/2005             0              0
## 6   7/6/2005 12/10/2005             2           1825
```

```
# Part b
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
pSmoking$admitdate <- pSmoking$admitdate %>%
  sapply(function(x) x[1]) %>% as.Date(format = c("%m/%d/%y"))

pSmoking$fdate <- pSmoking$fdate %>%
  sapply(function(x) x[1]) %>% as.Date(format = c("%m/%d/%y"))

head(pSmoking)
```

```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26             1
## 2 113  14       1   2  41      1    4          2           27             1
## 3  39   5       1   1  25      0    4          2           12             1
## 4  80  16       1   1  54      1    4          1           39             1
## 5  87   0       1   1  45      1    4          2           30             1
## 6  29 157       0   1  43      1    2          1           30             1
##    admitdate      fdate priorAttempts longestNoSmoke
## 1 2020-11-20 2020-12-31             0              0
## 2 2020-06-16 2020-06-30             3             90
## 3 2020-05-09 2020-05-14             3             21
## 4 2020-10-26 2020-11-11             0              0
## 5 2020-09-27 2020-09-27             0              0
## 6 2020-07-06 2020-12-10             2           1825
```

```r
# Part c
pSmoking$admitdateAlph <- format(pSmoking$admitdate, '%y%b%d')

pSmoking$fdateAlph <- format(pSmoking$fdate, '%y%b%d')

head(pSmoking)
```

```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26            1
## 2 113  14       1   2  41      1    4          2           27            1
## 3  39   5       1   1  25      0    4          2           12            1
## 4  80  16       1   1  54      1    4          1           39            1
## 5  87   0       1   1  45      1    4          2           30            1
## 6  29 157       0   1  43      1    2          1           30            1
##    admitdate      fdate priorAttempts longestNoSmoke admitdateAlph fdateAlph
## 1 2020-11-20 2020-12-31             0              0       20Nov20   20Dec31
## 2 2020-06-16 2020-06-30             3             90       20Jun16   20Jun30
## 3 2020-05-09 2020-05-14             3             21       20May09   20May14
## 4 2020-10-26 2020-11-11             0              0       20Oct26   20Nov11
## 5 2020-09-27 2020-09-27             0              0       20Sep27   20Sep27
## 6 2020-07-06 2020-12-10             2           1825       20Jul06   20Dec10
```

```r
# Part d
pSmoking$time <- difftime(pSmoking$fdate , pSmoking$admitdate, units = 'days') %>% as.numeric()

head(pSmoking)
```

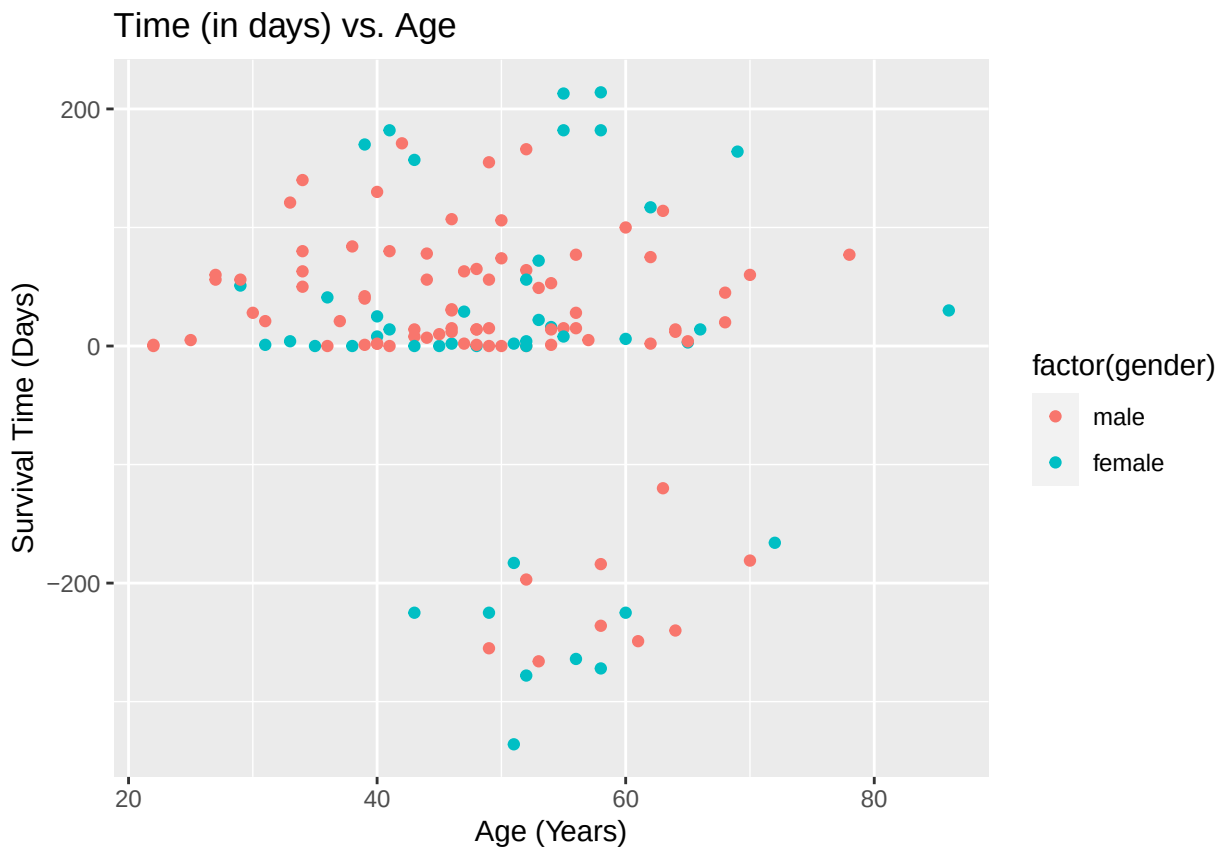
```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26            1
## 2 113  14       1   2  41      1    4          2           27            1
## 3  39   5       1   1  25      0    4          2           12            1
## 4  80  16       1   1  54      1    4          1           39            1
## 5  87   0       1   1  45      1    4          2           30            1
## 6  29 157       0   1  43      1    2          1           30            1
##    admitdate      fdate priorAttempts longestNoSmoke admitdateAlph fdateAlph
## 1 2020-11-20 2020-12-31             0              0       20Nov20   20Dec31
## 2 2020-06-16 2020-06-30             3             90       20Jun16   20Jun30
## 3 2020-05-09 2020-05-14             3             21       20May09   20May14
## 4 2020-10-26 2020-11-11             0              0       20Oct26   20Nov11
## 5 2020-09-27 2020-09-27             0              0       20Sep27   20Sep27
## 6 2020-07-06 2020-12-10             2           1825       20Jul06   20Dec10
##   time
## 1   41
## 2   14
## 3    5
## 4   16
## 5    0
## 6  157
```

```r
# Part e
head(pSmoking, 10)
```

```
##     id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1   21  41       0   2  36      1    4          1           26            1
## 2  113  14       1   2  41      1    4          2           27            1
## 3   39   5       1   1  25      0    4          2           12            1
## 4   80  16       1   1  54      1    4          1           39            1
## 5   87   0       1   1  45      1    4          2           30            1
## 6   29 157       0   1  43      1    2          1           30            1
## 7   16  14       1   2  66      1    1          3           54            1
## 8   35  77       1   2  78      0    1          2           56            2
## 9   54   2       1   2  40      0    1          1           25            1
## 10  70   0       1   2  38      1    1          1           23            2
##     admitdate      fdate priorAttempts longestNoSmoke admitdateAlph fdateAlph
## 1  2020-11-20 2020-12-31             0              0       20Nov20   20Dec31
## 2  2020-06-16 2020-06-30             3             90       20Jun16   20Jun30
## 3  2020-05-09 2020-05-14             3             21       20May09   20May14
## 4  2020-10-26 2020-11-11             0              0       20Oct26   20Nov11
## 5  2020-09-27 2020-09-27             0              0       20Sep27   20Sep27
## 6  2020-07-06 2020-12-10             2           1825       20Jul06   20Dec10
## 7  2020-11-03 2020-11-17             0              0       20Nov03   20Nov17
## 8  2020-03-26 2020-06-11            10             15       20Mar26   20Jun11
## 9  2020-10-04 2020-10-06             4              7       20Oct04   20Oct06
## 10 2020-11-26 2020-11-26            10             90       20Nov26   20Nov26
##    time
## 1    41
## 2    14
## 3     5
## 4    16
## 5     0
## 6   157
## 7    14
## 8    77
## 9     2
## 10    0
```

```r
# Part f
# Not using SAS therefore cannot use functions specified in homework
library(ggplot2)

ggplot(data = pSmoking) + geom_point(aes(x = age, y = time, color = factor(gender))) +
labs(title = 'Time (in days) vs. Age') +
xlab('Age (Years)') + ylab('Survival Time (Days)') +
scale_color_discrete(labels = c('male', 'female'))
```


Time (in days) vs. Age
Survival Time (Days)
200
0
−200
20
40
60
80
Age (Years)
factor(gender)
male
female