BIOST537_HW2

Joanna

1/27/2021

```
options(digits = 4)
knitr::opts_chunk$set(echo = TRUE)
```

```
# Load relevant packages
library(foreign)
library(survival)
library(flexsurv)
library(survMisc)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

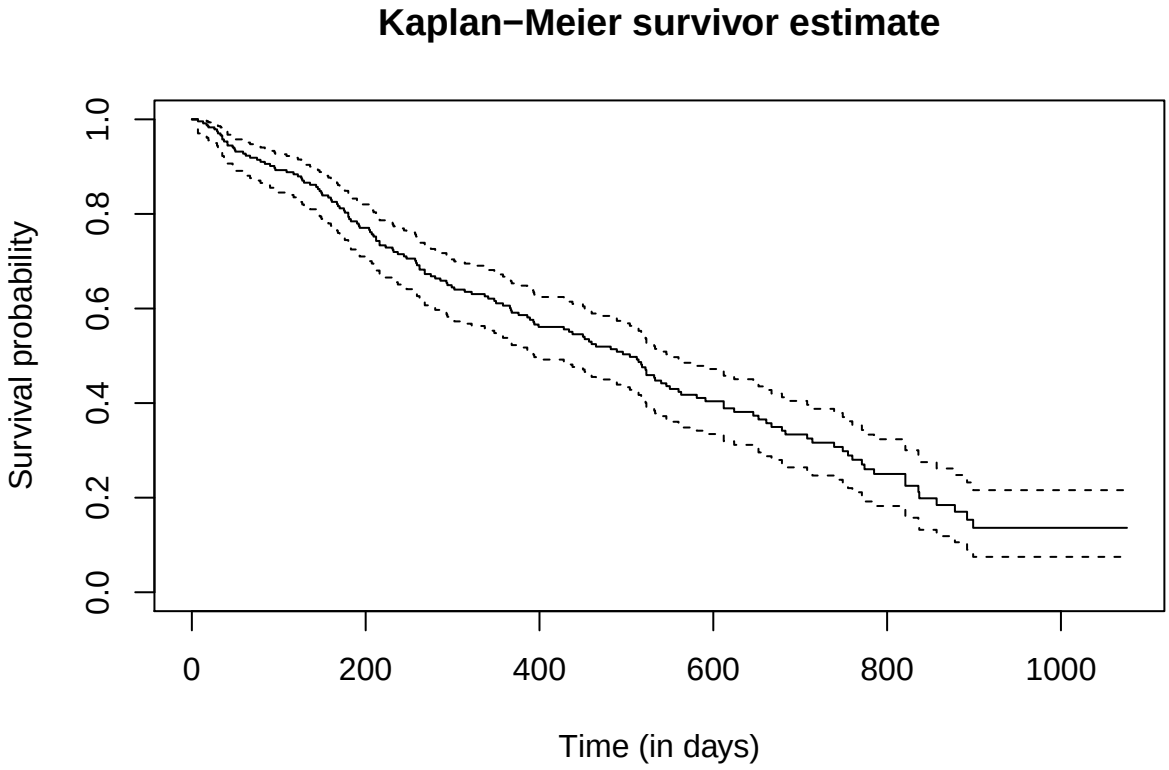
```
## The following object is masked from 'package:survMisc':
##
##     autoplot
```

```
## Loading required package: ggpubr
```

```
# read in data
source("/Users/ziyuxiao/Desktop/UW/Winter2021/BIOST537/Discussion/code/getmedianres.R")
data <- read.csv("/Users/ziyuxiao/Desktop/UW/Winter2021/BIOST537/HW data/addicts.csv")
```

Problem 2(a)

```
s.data <- with(data, Surv(time, event  == 1))
km.data <- survfit(s.data~1, conf.type="log-log")
plot(km.data,main="Kaplan-Meier survivor estimate",
     ylab="Survival probability",xlab="Time (in days)")
```




```
summary(km.data,times = 365)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365    122      87    0.606  0.0331        0.538        0.667
```

(a) The estimated probability that no exit will occur by one year is 0.606, and 95%CI is (0.538, 0.667).

Problem 2(b)

```
# by hand
# estimate is 504
summary(km.data,times = 500:520)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   500     92     107    0.503  0.0346        0.434        0.569
##   501     92       0    0.503  0.0346        0.434        0.569
##   502     92       0    0.503  0.0346        0.434        0.569
##   503     92       0    0.503  0.0346        0.434        0.569
##   504     92       1    0.498  0.0346        0.428        0.563
##   505     91       0    0.498  0.0346        0.428        0.563
##   506     91       0    0.498  0.0346        0.428        0.563
##   507     91       0    0.498  0.0346        0.428        0.563
##   508     91       0    0.498  0.0346        0.428        0.563
##   509     91       0    0.498  0.0346        0.428        0.563
##   510     91       0    0.498  0.0346        0.428        0.563
##   511     91       0    0.498  0.0346        0.428        0.563
```

```
##   512     91       1    0.492  0.0347        0.423        0.558
##   513     90       0    0.492  0.0347        0.423        0.558
##   514     90       1    0.487  0.0347        0.417        0.553
##   515     89       0    0.487  0.0347        0.417        0.553
##   516     89       0    0.487  0.0347        0.417        0.553
##   517     89       1    0.481  0.0348        0.412        0.547
##   518     87       1    0.476  0.0348        0.406        0.542
##   519     86       0    0.476  0.0348        0.406        0.542
##   520     86       0    0.476  0.0348        0.406        0.542
```

```
# lower 95%CI is 394
summary(km.data,times = 393:400)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   393    115      94    0.571  0.0337        0.502        0.634
##   394    114       1    0.566  0.0338        0.497        0.629
##   395    112       0    0.566  0.0338        0.497        0.629
##   396    112       0    0.566  0.0338        0.497        0.629
##   397    112       0    0.566  0.0338        0.497        0.629
##   398    112       0    0.566  0.0338        0.497        0.629
##   399    112       1    0.561  0.0339        0.492        0.624
##   400    111       0    0.561  0.0339        0.492        0.624
```

```
# upper 95%CI is 550
summary(km.data,times = 540:550)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   540     77     118    0.442   0.035        0.372        0.509
##   541     76       0    0.442   0.035        0.372        0.509
##   542     75       0    0.442   0.035        0.372        0.509
##   543     75       0    0.442   0.035        0.372        0.509
##   544     74       0    0.442   0.035        0.372        0.509
##   545     74       0    0.442   0.035        0.372        0.509
##   546     74       1    0.436   0.035        0.366        0.503
##   547     73       0    0.436   0.035        0.366        0.503
##   548     73       0    0.436   0.035        0.366        0.503
##   549     73       0    0.436   0.035        0.366        0.503
##   550     73       1    0.430   0.035        0.361        0.497
```

```
# second way
print(km.data)
```

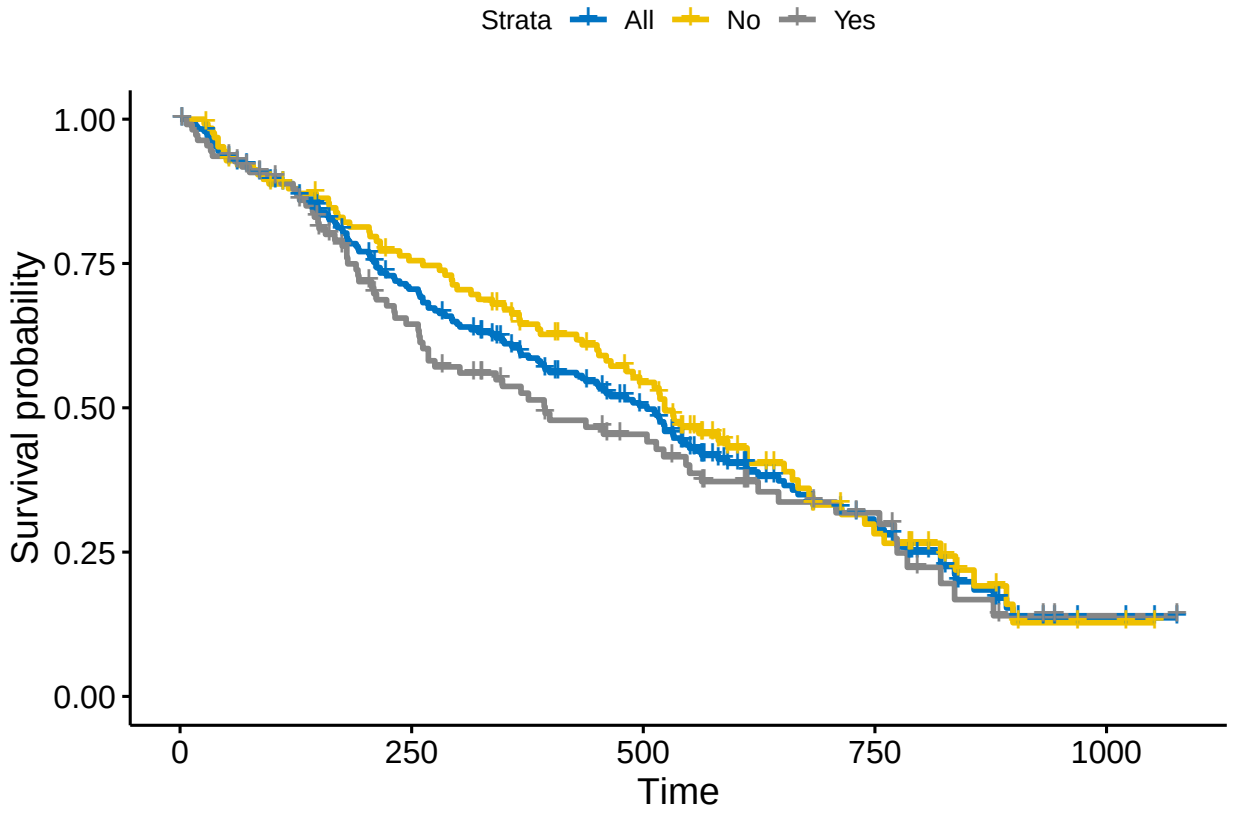
```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##       n  events  median 0.95LCL 0.95UCL
##     238     150     504     394     550
```

(b)(i) The estimated median time until exit from maintenance is 504 days. 95%CI is (394, 550). Construct the interval that includes all values of t such that the test of H0: S(t)=0.5 is not rejected. Therefore, we can find the lower and higher bound from the results.

(b)(ii) Using the print command. We can get the same results as above.

Problem 2(c)

```
km.data.by.incar <- survfit(s.data~prison,data=data,conf.type="log-log")
fit <- list(km.data,km.data.by.incar )
ggsurvplot_combine(fit,data,pval = TRUE,palette = "jco",
                   risk.table = FALSE,legend.labs = c("All","No","Yes"))
```




```
summary(km.data.by.incar,times = 240)
```

```
## Call: survfit(formula = s.data ~ prison, data = data, conf.type = "log-log")
##
##                 prison=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     240.0000      91.0000      29.0000       0.7634       0.0384       0.6776
## upper 95% CI
##       0.8292
##
##                 prison=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     240.0000      62.0000      35.0000       0.6555       0.0475       0.5537
## upper 95% CI
##       0.7395
```

```
sterr <- sqrt(0.0384^2+0.0475^2)
est <- 0.7634-0.6555
est/sterr
```

```
## [1] 1.767
```

(c)(ii) Since the test statistics is equal to $1.77 < 1.96$, we fail to reject the null hypothesis that the probability

that no exit occurred by 8 months is not different from each other in the two groups.

```
survdiff(s.data~prison,data = data)
```

```
## Call:
## survdiff(formula = s.data ~ prison, data = data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## prison=0 127       81     87.8     0.519      1.26
## prison=1 111       69     62.2     0.732      1.26
##
##  Chisq= 1.3  on 1 degrees of freedom, p= 0.3
```

(c)(iii) Since p is equal to $0.3 > 0.05$, we fail to reject the null hypothesis that the distribution of time until exit from maintenance is not different from each other.

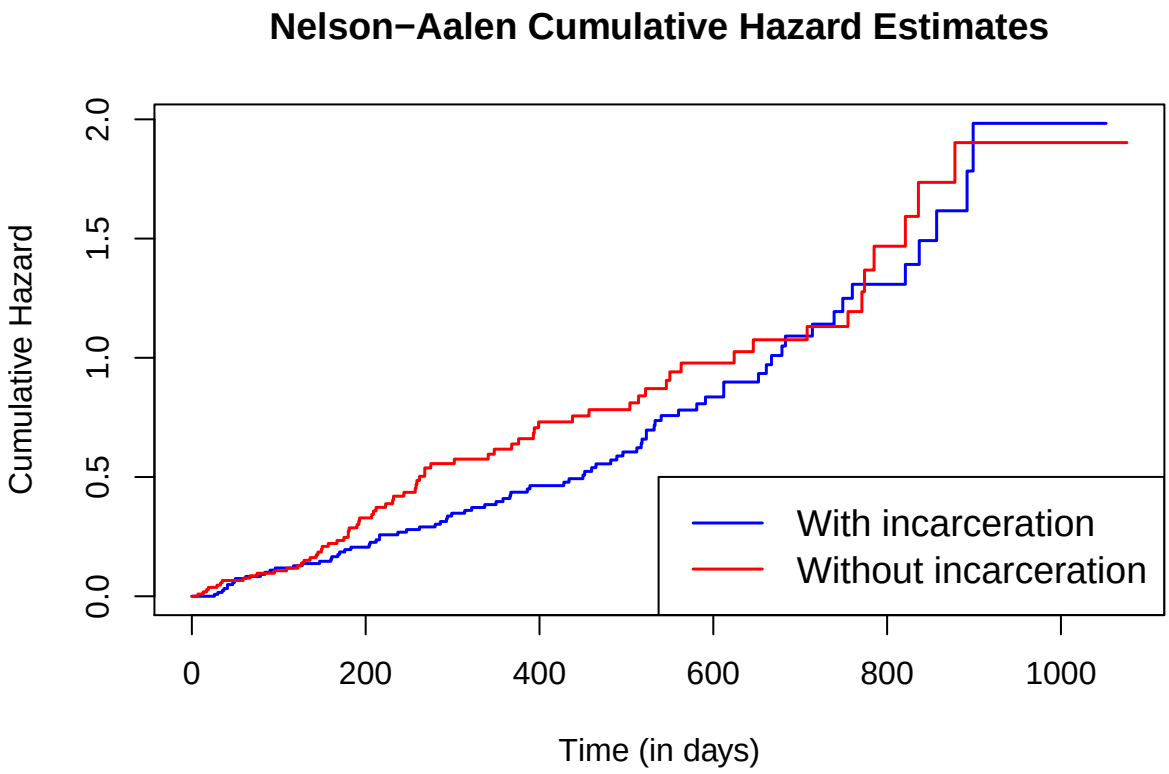
```
comp(ten(km.data.by.incar))$tests$lrTests
```

```
##                    Q      Var     Z pNorm
## 1           6.75e+00 3.62e+01 1.122     4
## n           1.43e+03 8.04e+05 1.600     1
## sqrtN       1.01e+02 4.86e+03 1.454     3
## S1          6.24e+00 1.69e+01 1.515     2
## S2          6.21e+00 1.67e+01 1.519     2
## FH_p=1_q=1  8.59e-01 1.31e+00 0.749     5
##              maxAbsZ      Var    Q pSupBr
## 1           1.07e+01 3.62e+01 1.78      5
## n           1.81e+03 8.04e+05 2.02      1
## sqrtN       1.39e+02 4.86e+03 1.99      4
## S1          8.26e+00 1.69e+01 2.01      3
## S2          8.21e+00 1.67e+01 2.01      2
## FH_p=1_q=1  1.78e+00 1.31e+00 1.55      6
```

```
## NULL
```

(c)(iv) By looking at the Z statistics for Wilcoxon-Gehan-Breslow test, which is equal to $1.6 < 1.96$, we therefore fail to reject the null hypothesis that the distribution of time until exit from maintenance is not different from each other.

```
plot(km.data.by.incar,fun="cumhaz",col=c("blue","red"),lwd=1.5,conf.int=FALSE,
     xlab = "Time (in days)",
     ylab = "Cumulative Hazard",
     main = "Nelson-Aalen Cumulative Hazard Estimates")
legend("bottomright",c("With incarceration","Without incarceration"),
       col = c("blue","red"),lwd =c(1.5,1.5),cex=1.2)
```

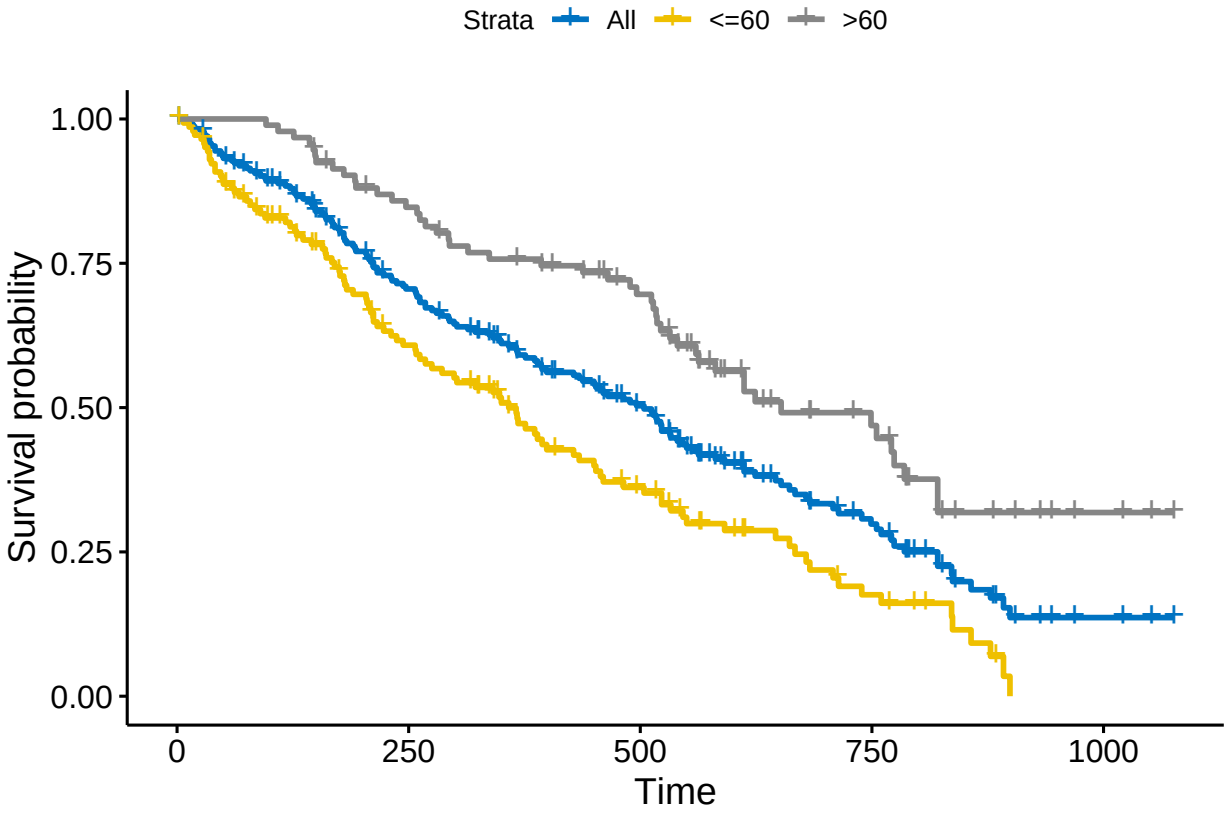


(c)(v) From the cumulative hazard estimates plot, the two groups cross over at some points. The logrank test has very little power against crossing hazard functions. We can use the weighted logrank test, such as the Wilcoxon-Gehan-Breslow test here if we consider the survival differences are meaningful at earlier times.

Problem 2(d)

```
data$dose_binary <- ifelse(data$dose>60,1,0)
km.data.by.dose <- survfit(s.data~dose_binary,data=data,conf.type="log-log")

fit.dose <- list(km.data,km.data.by.dose )
ggsurvplot_combine(fit.dose,data,pval = TRUE,palette = "jco",
                   risk.table = FALSE,legend.labs = c("All","<=60",">60"))
```




```
summary(km.data.by.dose,times = 240)
```

```
## Call: survfit(formula = s.data ~ dose_binary, data = data, conf.type = "log-log")
##
##                 dose_binary=0
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     240.0000      76.0000      51.0000       0.6162       0.0425       0.5272
## upper 95% CI
##       0.6934
##
##                 dose_binary=1
##         time       n.risk      n.event     survival      std.err lower 95% CI
##     240.0000      77.0000      13.0000       0.8582       0.0365       0.7682
## upper 95% CI
##       0.9151
```

```
sterr <- sqrt(0.0425^2+0.0365^2)
est <- 0.6162-0.8582
est/sterr
```

```
## [1] -4.32
```

(d)(ii) Since the test statistics is equal to -4.32 < -1.96, we reject the null hypothesis that the probability that no exit occurred by 8 months is not different from each other in the two groups.

```
survdiff(s.data~dose_binary,data = data)
```

```
## Call:
## survdiff(formula = s.data ~ dose_binary, data = data)
##
```

```
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## dose_binary=0 145      102       71      13.6      26.5
## dose_binary=1  93       48       79      12.2      26.5
##
##  Chisq= 26.5  on 1 degrees of freedom, p= 3e-07
```

(d)(iii) Since p is equal to 3e-07 < 0.05, we reject the null hypothesis that the distribution of time until exit from maintenance is not different from each other.

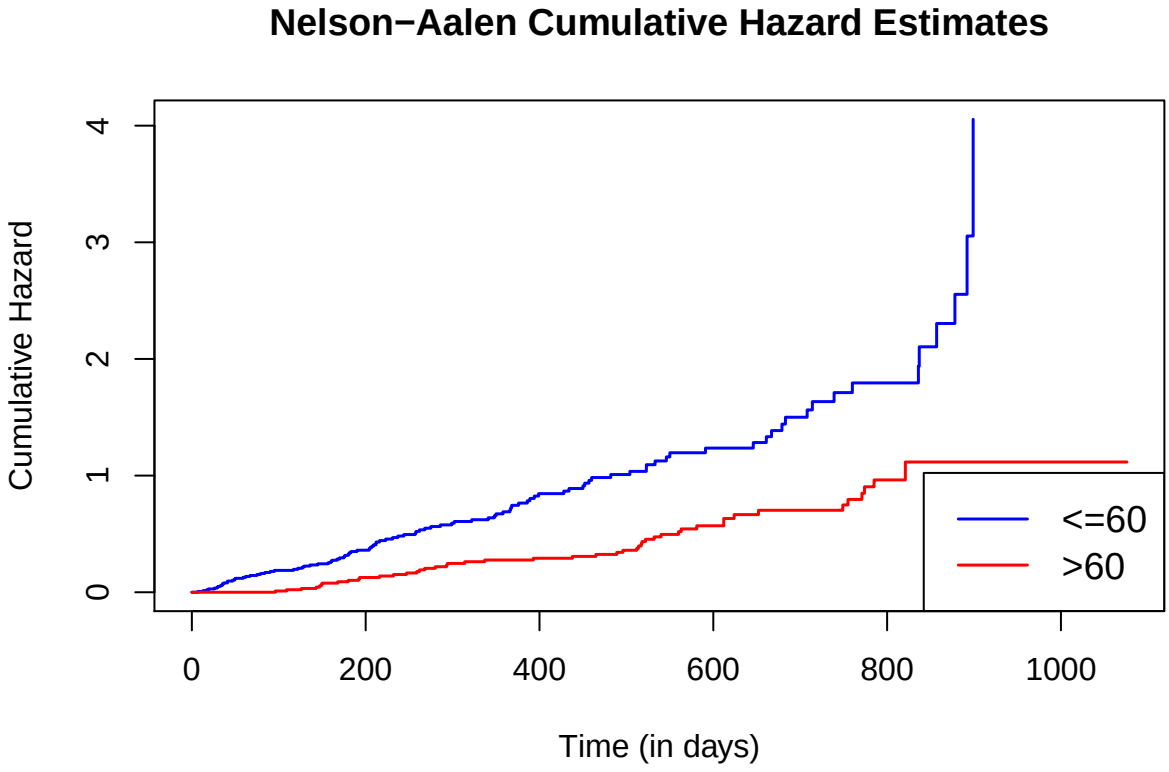
```
comp(ten(survfit(s.data~dose_binary,data = data)))$tests$lrTests
```

```
##                    Q       Var     Z pNorm
## 1              -31.05     36.34 -5.15     1
## n            -4459.00 810346.87 -4.95     5
## sqrtN         -354.13   4922.56 -5.05     4
## S1             -20.91     17.11 -5.06     3
## S2             -20.73     16.88 -5.05     4
## FH_p=1_q=1      -4.76      1.34 -4.11     2
##              maxAbsZ      Var    Q pSupBr
## 1           3.11e+01 3.63e+01 5.15      2
## n           4.46e+03 8.10e+05 4.95      1
## sqrtN       3.54e+02 4.92e+03 5.05      5
## S1          2.09e+01 1.71e+01 5.06      4
## S2          2.07e+01 1.69e+01 5.05      5
## FH_p=1_q=1 4.76e+00 1.34e+00 4.11      3
```

```
## NULL
```

(d)(iv) By looking at the Z statistics for Wilcoxon-Gehan-Breslow test, which is equal to -4.95 $<$ -1.96, we therefore reject the null hypothesis that the distribution of time until exit from maintenance is not different from each other.

```
plot(km.data.by.dose,fun="cumhaz",col=c("blue","red"),lwd=1.5,conf.int=FALSE,
     xlab = "Time (in days)",
     ylab = "Cumulative Hazard",
     main = "Nelson-Aalen Cumulative Hazard Estimates")
legend("bottomright",c("<=60",">60"),
       col = c("blue","red"),lwd =c(1.5,1.5),cex=1.2)
```



(d)(v)

From the cumulative hazard estimates plot, the two groups do not cross over all the time. We can directly use the logrank test here. It has enough power. And both the two tests have the similar results for the hypothesis testing.

Problem 2(e)

```
survdiff(s.data~prison+strata(clinic),data=data)
```

```
## Call:
## survdiff(formula = s.data ~ prison + strata(clinic), data = data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## prison=0 127       81     92.7      1.48      4.04
## prison=1 111       69     57.3      2.40      4.04
##
##  Chisq= 4  on 1 degrees of freedom, p= 0.04
```

```
survdiff(s.data~prison,data=data)
```

```
## Call:
## survdiff(formula = s.data ~ prison, data = data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## prison=0 127       81     87.8     0.519      1.26
## prison=1 111       69     62.2     0.732      1.26
##
##  Chisq= 1.3  on 1 degrees of freedom, p= 0.3
```

The null hypothesis is that for either clinic 1 or clinic 2 group, the hazard is the same in prison group as in non-prison group.

The alternative hypothesis is that for either clinic 1 or clinic 2 group, the hazard is not the same in prison

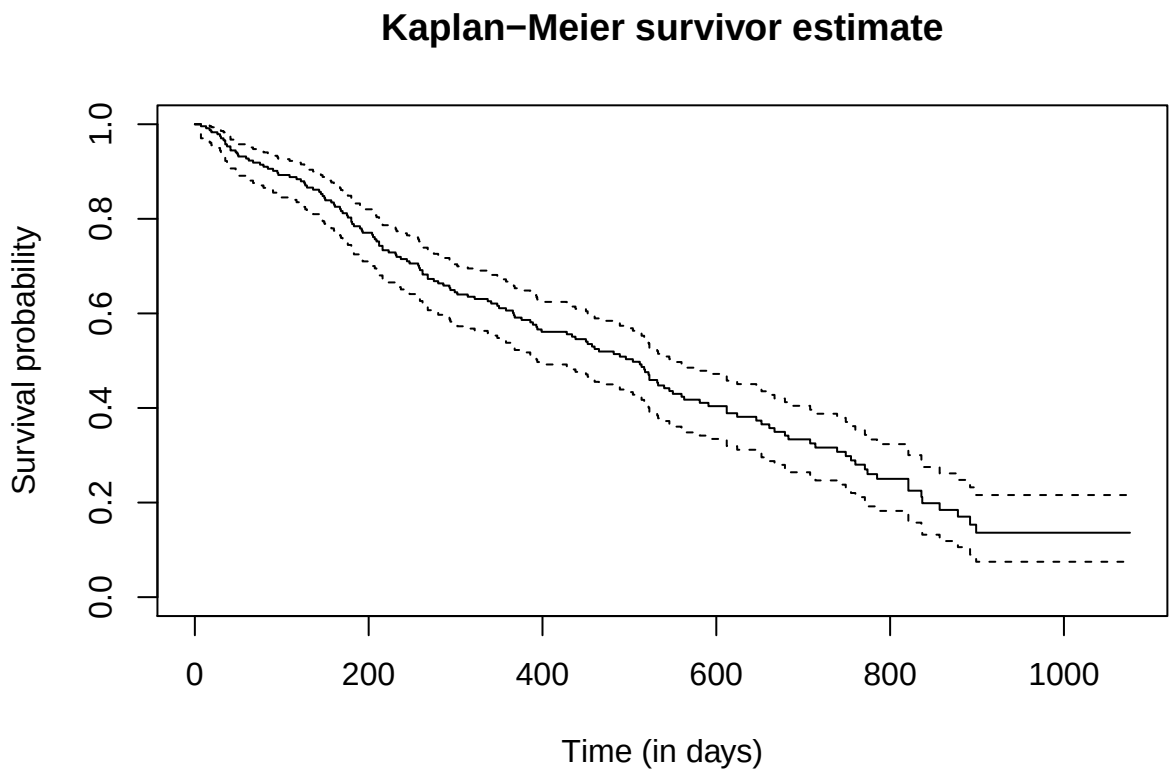
group as in non-prison group at some time t.

Since p is equal to $0.04 < 0.05$. We reject the null hypothesis that the time until exit from maintenance is not different by history of previous incarceration adjusting for clinic membership.

In the standard logrank test, since p is equal to $0.3 > 0.05$. We fail to reject the null hypothesis that the time until exit from maintenance is not different by history of previous incarceration.

Problem 2(f)

```
plot(km.data,main="Kaplan-Meier survivor estimate",
     ylab="Survival probability",xlab="Time (in days)")
```




```
summary(km.data,times = 120)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   120    199      27    0.884   0.021        0.835        0.919
```

```
0.884/2 #0.442
```

```
## [1] 0.442
```

```
summary(km.data,times = 530:550)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   530     83     115    0.459  0.0349        0.390        0.526
##   531     83       0    0.459  0.0349        0.390        0.526
##   532     80       1    0.453  0.0349        0.384        0.520
```

```
##   533     78       1    0.448  0.0349        0.378        0.515
##   534     77       0    0.448  0.0349        0.378        0.515
##   535     77       0    0.448  0.0349        0.378        0.515
##   536     77       0    0.448  0.0349        0.378        0.515
##   537     77       0    0.448  0.0349        0.378        0.515
##   538     77       0    0.448  0.0349        0.378        0.515
##   539     77       0    0.448  0.0349        0.378        0.515
##   540     77       1    0.442  0.0350        0.372        0.509
##   541     76       0    0.442  0.0350        0.372        0.509
##   542     75       0    0.442  0.0350        0.372        0.509
##   543     75       0    0.442  0.0350        0.372        0.509
##   544     74       0    0.442  0.0350        0.372        0.509
##   545     74       0    0.442  0.0350        0.372        0.509
##   546     74       1    0.436  0.0350        0.366        0.503
##   547     73       0    0.436  0.0350        0.366        0.503
##   548     73       0    0.436  0.0350        0.366        0.503
##   549     73       0    0.436  0.0350        0.366        0.503
##   550     73       1    0.430  0.0350        0.361        0.497
```

```
# time is 540 days
# median residual time
540-120
```

```
## [1] 420
```

```
summary(km.data,times = 240)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   240    153      64    0.715  0.0302        0.651        0.769
```

```
0.715/2 #0.3575
```

```
## [1] 0.3575
```

```
summary(km.data,times = 660:670)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   660     46     129    0.365  0.0359        0.296        0.435
##   661     46       1    0.357  0.0359        0.288        0.428
##   662     45       0    0.357  0.0359        0.288        0.428
##   663     45       0    0.357  0.0359        0.288        0.428
##   664     45       0    0.357  0.0359        0.288        0.428
##   665     45       0    0.357  0.0359        0.288        0.428
##   666     45       0    0.357  0.0359        0.288        0.428
##   667     45       1    0.350  0.0360        0.280        0.420
##   668     44       0    0.350  0.0360        0.280        0.420
##   669     44       0    0.350  0.0360        0.280        0.420
##   670     44       0    0.350  0.0360        0.280        0.420
```

```
# time is 667
# median residual time
667-240
```

```
## [1] 427
```

```r
summary(km.data,times = 365)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   365    122      87    0.606  0.0331        0.538        0.667
```

```r
0.606/2 #0.303
```

```
## [1] 0.303
```

```r
summary(km.data,times = 730:750)
```

```
## Call: survfit(formula = s.data ~ 1, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   730     36     135    0.316  0.0362        0.247        0.388
##   731     35       0    0.316  0.0362        0.247        0.388
##   732     35       0    0.316  0.0362        0.247        0.388
##   733     35       0    0.316  0.0362        0.247        0.388
##   734     35       0    0.316  0.0362        0.247        0.388
##   735     35       0    0.316  0.0362        0.247        0.388
##   736     35       0    0.316  0.0362        0.247        0.388
##   737     35       0    0.316  0.0362        0.247        0.388
##   738     35       0    0.316  0.0362        0.247        0.388
##   739     35       1    0.307  0.0363        0.238        0.379
##   740     34       0    0.307  0.0363        0.238        0.379
##   741     34       0    0.307  0.0363        0.238        0.379
##   742     34       0    0.307  0.0363        0.238        0.379
##   743     34       0    0.307  0.0363        0.238        0.379
##   744     34       0    0.307  0.0363        0.238        0.379
##   745     34       0    0.307  0.0363        0.238        0.379
##   746     34       0    0.307  0.0363        0.238        0.379
##   747     34       0    0.307  0.0363        0.238        0.379
##   748     34       0    0.307  0.0363        0.238        0.379
##   749     34       1    0.298  0.0363        0.229        0.370
##   750     33       0    0.298  0.0363        0.229        0.370
```

```r
# time is 749
# median residual time
749-365
```

```
## [1] 384
```

The estimated median residual time is **420**, **427**, **384** respectively.

```r
m1 <- getmedianres(survobj = s.data, times = 120, confint = TRUE)
m2 <- getmedianres(survobj = s.data, times = 240, confint = TRUE)
m3 <- getmedianres(survobj = s.data, times = 365, confint = TRUE)
```

The estimated median residual time until exit from maintenance at 120 days is 420. The estimated median residual time until exit from maintenance at 240 days is 427. The estimated median residual time until exit from maintenance at 365 days is 384. The 95% CI are [369, 526],[323, 515], [296, 456] respectively.