BIOST 537 HOMEWORK2

Ivy Zhang

1/26/2022

Problem 1

Part a

```
# Problem 1
library(dplyr)
library(knitr)
main_tab = data.frame(time_t = c(0,8,11,14,20,23,28,31,33,45,49,152),
                      n_k = c(12,12,11,10,9,8,7,6,5,4,2,1),
                      d_k = c(0,1,1,0,1,1,0,1,1,1,1,0))
main_tab = main_tab%>%mutate(d_n = round(d_k/n_k,3),
                             d_n_re = round(1-d_k/n_k,3))
main_tab = main_tab %>% mutate(s = round(cumprod(d_n_re),3),
                               h = cumsum(d_n))
colnames(main_tab) = c("time t","# at risk(n)","# events(d)","d/n","1 - d/n","S(t)","H(t)")
kable(main_tab)
```

time t	# at risk(n)	# events(d)	d/n	1 - d/n	S(t)	H(t)
0	12	0	0.000	1.000	1.000	0.000
8	12	1	0.083	0.917	0.917	0.083
11	11	1	0.091	0.909	0.834	0.174
14	10	0	0.000	1.000	0.834	0.174
20	9	1	0.111	0.889	0.741	0.285
23	8	1	0.125	0.875	0.648	0.410
28	7	0	0.000	1.000	0.648	0.410
31	6	1	0.167	0.833	0.540	0.577
33	5	1	0.200	0.800	0.432	0.777
45	4	1	0.250	0.750	0.324	1.027
49	2	1	0.500	0.500	0.162	1.527
152	1	0	0.000	1.000	0.162	1.527

```
control_tab = data.frame(time_t = c(0,3,5,7,10,12,16,25,27,30,38,44,48),
                         n_k = c(13,13,12,11,9,8,7,6,5,4,3,2,1),
                         d_k = c(0,1,1,2,0,1,0,1,1,1,1,1,1))
control_tab = control_tab%>%mutate(d_n = round(d_k/n_k,3),
                             d_n_re = round(1-d_k/n_k,3))
control_tab = control_tab %>% mutate(s = round(cumprod(d_n_re),3),
                               h = cumsum(d_n))
colnames(control_tab) = c("time t","# at risk(n)","# events(d)","d/n","1 - d/n","S(t)","H(t)")
```

```
kable(control_tab)
```

time t	# at risk(n)	# events(d)	d/n	1 - d/n	S(t)	H(t)
0	13	0	0.000	1.000	1.000	0.000
3	13	1	0.077	0.923	0.923	0.077
5	12	1	0.083	0.917	0.846	0.160
7	11	2	0.182	0.818	0.692	0.342
10	9	0	0.000	1.000	0.692	0.342
12	8	1	0.125	0.875	0.606	0.467
16	7	0	0.000	1.000	0.606	0.467
25	6	1	0.167	0.833	0.505	0.634
27	5	1	0.200	0.800	0.404	0.834
30	4	1	0.250	0.750	0.303	1.084
38	3	1	0.333	0.667	0.202	1.417
44	2	1	0.500	0.500	0.101	1.917
48	1	1	1.000	0.000	0.000	2.917

Part(b)

```
main_close = max(main_tab$`time t`[main_tab$`time t`<=36])
(main_36 = main_tab[main_tab$`time t` == main_close,"S(t)"])
```

```
## [1] 0.432
```

```
control_close = max(control_tab$`time t`[control_tab$`time t`<=36])
(control_36 = control_tab[control_tab$`time t` == control_close,"S(t)"])
```

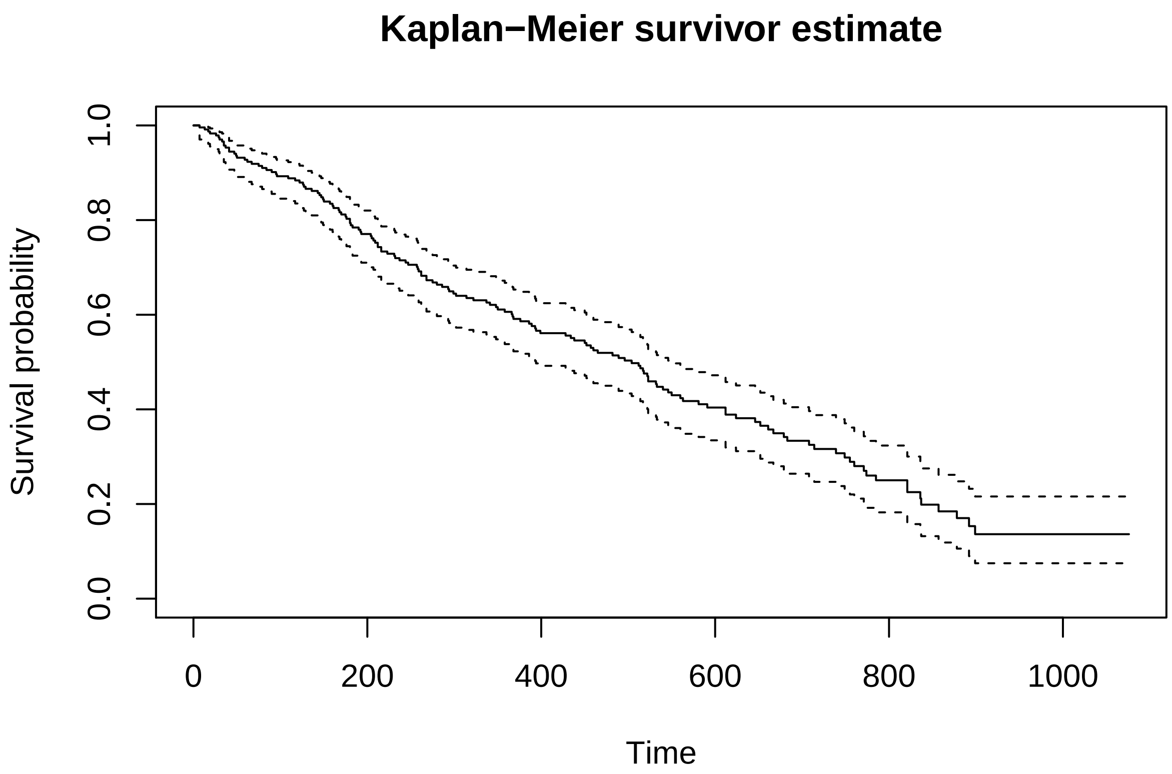
```
## [1] 0.303
```

We estimate that the probability that no relapse will occur by 36 months is 0.432 for the maintenance group, and 0.303 for the contro group.

Problem 2

part(a) Plot the Kaplan-Meier estimator of the survival function of the time until exit from maintenance along with pointwise 95% confidence intervals.

```
#Problem 2
library(readr)
library(survival)
library(foreign)
library(flexsurv)
library(msm)
addicts <- read_csv("~/Desktop/R hw/addicts.csv")
s.addicts = with(addicts, Surv(time, event))
km.addicts = survfit(s.addicts~1, conf.type = "log-log")
plot(km.addicts, main = "Kaplan-Meier survivor estimate", ylab = "Survival probability",
     xlab = "Time", cex = 0.5)
```



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

```
## Call: survfit(formula = s.addicts ~ 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
```

Based on the calculation using survfit function, the estimated probability that no exit occur before one year is 0.606 (95% CI: 0.538, 0.667).

part(b) Provide the estimated median time until exit from maintenance and associated 95% confidence interval by:

We first try to look at the Kaplan-Meier graph only to have an estimated median time and its 95% confidence interval.

```
#b
median_time = min(km.addicts$time[km.addicts$surv<=0.5], na.rm = TRUE)
median_time_low = min(km.addicts$time[km.addicts$lower<=0.5],na.rm = TRUE)
median_time_up = min(km.addicts$time[km.addicts$upper<0.5],na.rm = TRUE)
(c(median_time, median_time_low, median_time_up))
```

```
## [1] 504 394 550
```

Our estimated median time is 504, and our estimated 95% confidence intervel for estimated median time is [394.0, 550.0]. It is a interval that includes all values of t such that the test of null hypothesis of S(t) = 0.5 is not rejected at the confidence level of 0.95 using Kaplan-Meier estimation. 504 is the smallest time such that its survival rate is smaller than 0.5.

Then we try to find the median estimate and confidence intervals provided by the survfit command.

```
km.addicts
```

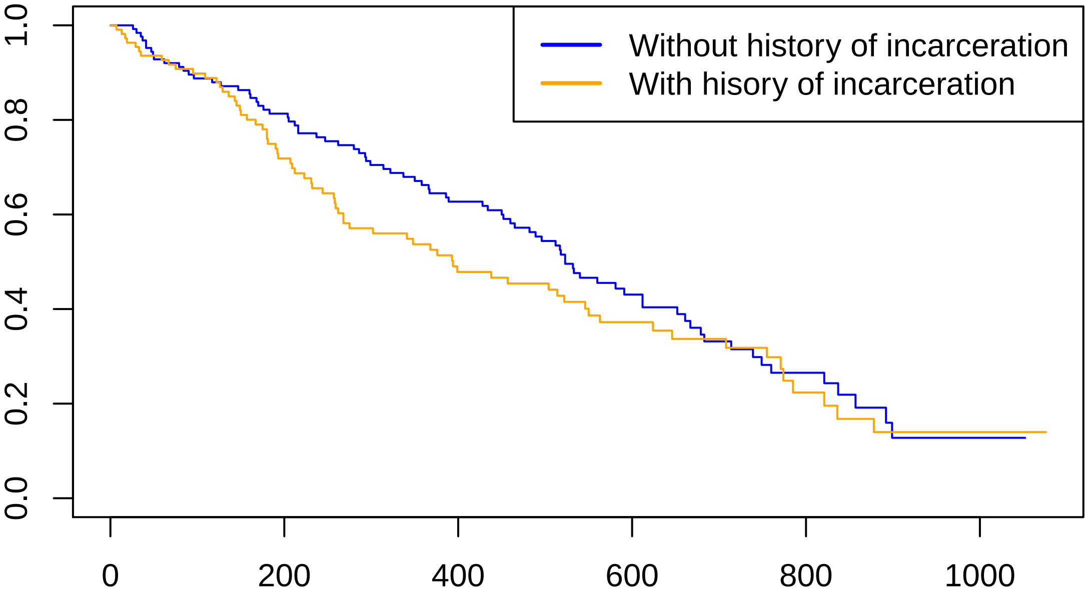
```
## Call: survfit(formula = s.addicts ~ 1, conf.type = "log-log")
##
##       n  events  median 0.95LCL 0.95UCL
##     238     150     504     394     550
```

Based on the survfit function, the estimated median time is also 504, with the same 95% confidence interval is [394,550].

part(c)

We first plotting the Kaplan-Meier estimator of the survival function of the time until exit from maintenance for patients with a history of incarceration and for patients without.

```
#c
km.addicts.prison = survfit(s.addicts~prison, data = addicts)
plot(km.addicts.prison, col = c("blue", "orange"))
legend("topright", c("Without history of incarceration","With hisory of incarceration"),
       col = c("blue","orange"), lwd = c(2, 2))
```



We then want to figure out does the probability that no exit occurred by 8 months differ significantly between these two groups. We will use Wald test to test it. For 8 months, we will transform it to 240 days by using a month of 30 days times 8.

```
s <- summary(km.addicts.prison, times = 240)
diff <- s$surv[1] - s$surv[2]
se <- sqrt(s$std.err[1]^2 + s$std.err[2]^2)
diff/se
```

```
## [1] 1.765927
```

```
(1-pnorm(abs(diff)/se)) * 2
```

```
## [1] 0.07740806
```

By using the wald test, we have a calculated a wald test statistic value of 1.766 and associated p-value of 0.077 which is larger than 0.05. Therefore, we do not have enough evidence from data to reject the null hypothesis that there is no difference of the probability that no exit occurred by 8 months between these two groups at the significance level of 0.05.

We then want to use the logrank test to see does the distribution of time until exit from maintenance differ significantly by history of incarceration.

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

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

Based on the calculation of the results of the logrank test, we can see we have a calculated chi-square statistics of 1.3 and associated p-value of 0.3 which is larger than 0.05. Therefore, we do not have enough evidence from the data to reject the null hypothesis that the distribution of time until exit from maintenance are same by history of incarceration at the significance level of 0.05.

Then we try to usethe Wilcoxon-Gehan-Breslow test to see does the distribution of time until exit from maintenance differ significantly by history of incarceration.

```
library(survMisc)
comp(ten(km.addicts.prison))$testes$lrTests
```

```
##                      Q        Var       Z pNorm
## 1           6.7504e+00 3.6213e+01 1.12176     5
## n           1.4340e+03 8.0365e+05 1.59962     1
## sqrtN       1.0139e+02 4.8628e+03 1.45397     4
## S1          6.2348e+00 1.6946e+01 1.51458     3
## S2          6.2107e+00 1.6714e+01 1.51912     2
## FH_p=1_q=1 8.5884e-01 1.3149e+00 0.74898     6
##                maxAbsZ        Var      Q pSupBr
## 1           1.0712e+01 3.6213e+01 1.7801      5
## n           1.8130e+03 8.0365e+05 2.0224      1
## sqrtN       1.3901e+02 4.8628e+03 1.9934      4
## S1          8.2577e+00 1.6945e+01 2.0060      3
## S2          8.2090e+00 1.6714e+01 2.0079      2
## FH_p=1_q=1 1.7753e+00 1.3149e+00 1.5483      6
```

```
## NULL
```

```
(1-pnorm(abs(1.59962)))*2
```

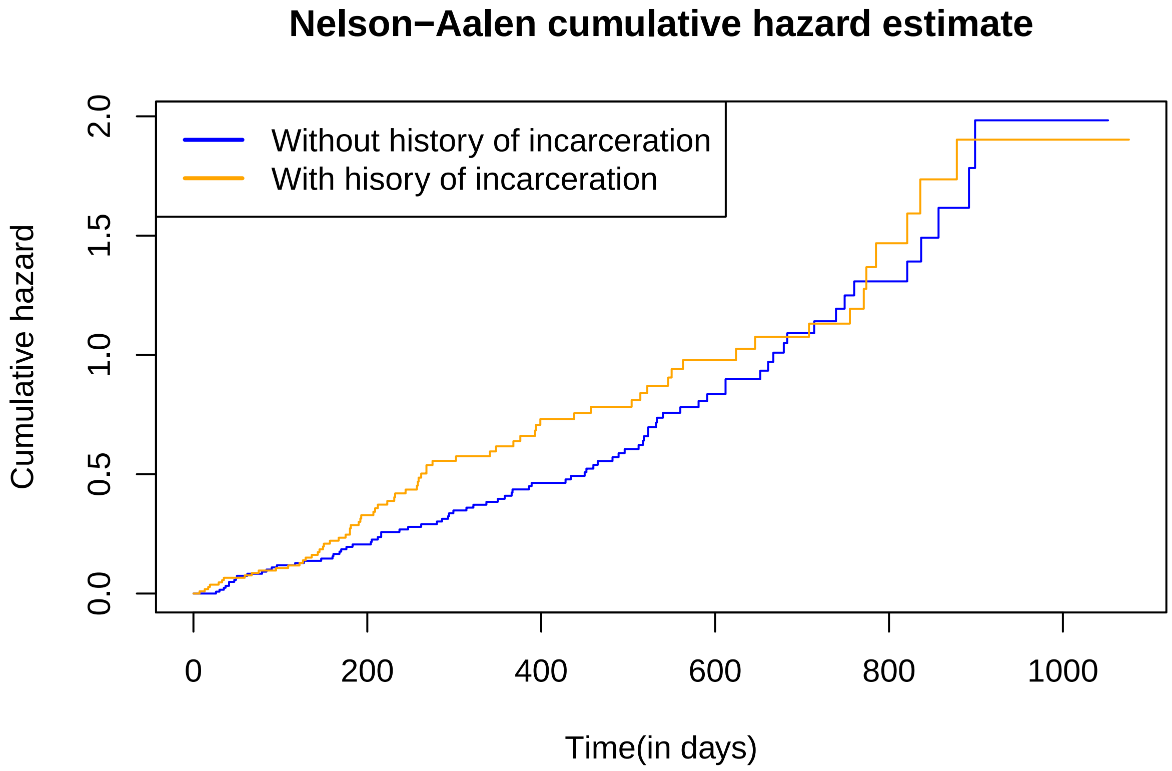
```
## [1] 0.1096829
```

Based on the calculation, we can see that the Wilcoxon-Gehan-Breslow test gave us a Z-value of 1.59962 which gave us a two-sided p-value that is 0.1096829. Therefore, based on the Wilcoxon-Gehan-Breslow test, we do not have evidence from the data to reject null hypothesis that there is no difference between the distribution of time until exit from maintenance in two groups with different history of incarceration.

Then we are going to plot the estimated hazard functions for patients with and without a history of incarceration.

```
plot(km.addicts.prison, col = c("blue", "orange"), fun = "cumhaz",xlab = "Time(in days)",
     main = "Nelson-Aalen cumulative hazard estimate", ylab = "Cumulative hazard")
```

```r
legend("topleft", c("Without history of incarceration","With hisory of incarceration"),
       col = c("blue","orange"), lwd = c(2, 2))
```

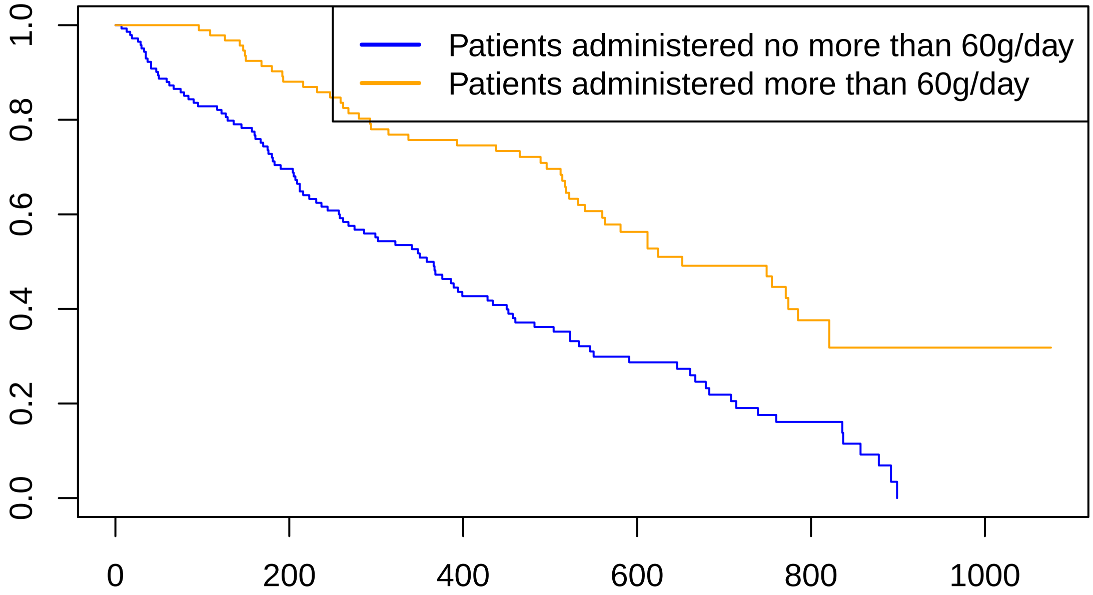


As we can see from the plot, the cumulative hazard distribution does not seem much differ and the hazard function are crossing. Therefore, it may lead the log-rank test has less power, and therefore give us a relatively large p-value. Wilcoxon-Gehan-Breslow tests gave us relatively larger test statistics and smaller p-value compared to standard log-rank test.It may because Wilcoxon-Gehan-Breslow may weight the earlier stage more than the log-rank test and the cumulative hazard function of the two group differ moreat the earlier stage. These two curves may not differ much enough to actually for both two tests to detetect the significantly difference.

part(d)

We first dictomize an indicator variable showing whether the participant is having more than 60mg of methadone and then plot the Kaplan-Meier estimator of the survival function of the time until exit from maintenance for patients in these two groups differed by this indicator.

```r
#d
addicts$do60 = addicts$dose>60
km.addicts.dose = survfit(s.addicts~do60, data = addicts)
plot(km.addicts.dose, col = c("blue", "orange"))
legend("topright", c("Patients administered no more than 60g/day",
                     "Patients administered more than 60g/day"),
       col = c("blue","orange"), lwd = c(2, 2))
```



We then want to figure out does the probability that no exit occurred by 8 months differ significantly between these two groups. We will use Wald test to test it. For 8 months, we will transform it to 243.33 days by using a month of 30 days times 8.

```
s <- summary(km.addicts.dose, times = 240)
diff <- s$surv[1] - s$surv[2]
se <- sqrt(s$std.err[1]^2 + s$std.err[2]^2)
diff/se
```

```
## [1] -4.31897
```

```
(1-pnorm(abs(diff)/se, lower.tail = T)) * 2
```

```
## [1] 1.56759e-05
```

By using the wald test, we have a wald test statistics value of -4.319 and associated calculated p-value of 1.568×10^{-5} which is much smaller than 0.05. Therefore, we do have enough evidence from data to reject the null hypothesis that there is no difference of the probability that no exit occurred by 8 months between these two groups at the significance level of 0.05.

We then want to use the logrank test to see does the distribution of time until exit from maintenance differ significantly by high dose of methadone.

```
survdiff(s.addicts~do60, data =addicts)
```

```
## Call:
## survdiff(formula = s.addicts ~ do60, data = addicts)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## do60=FALSE 145      102       71      13.6      26.5
## do60=TRUE   93       48       79      12.2      26.5
##
##  Chisq= 26.5  on 1 degrees of freedom, p= 3e-07
```

Based on the calculation of the results of the logrank test, we can see we have a calculated chi-square statistics of 26.5 and associated p-value of 3×10^{-7} which is much smaller than 0.05. Therefore, we do have enough evidence from the data to reject the null hypothesis that the distribution of time until exit from maintenance are same by history of incarceration at the significance level of 0.05.

Then we try to usethe Wilcoxon-Gehan-Breslow test to see does the distribution of time until exit from maintenance differ significantly by history of incarceration.

```
comp(ten(km.addicts.dose))$testes$lrTests
```

```
##                      Q         Var       Z pNorm
## 1             -31.0457     36.3422 -5.1499     1
## n           -4459.0000 810346.8693 -4.9534     6
## sqrtN        -354.1274   4922.5602 -5.0474     4
## S1            -20.9117     17.1100 -5.0555     3
## S2            -20.7304     16.8777 -5.0460     5
## FH_p=1_q=1     -4.7555      1.3376 -4.1119     2
##                maxAbsZ        Var      Q pSupBr
## 1          3.1046e+01 3.6342e+01 5.1499      2
## n          4.4590e+03 8.1035e+05 4.9534      1
## sqrtN      3.5413e+02 4.9226e+03 5.0474      5
## S1         2.0912e+01 1.7110e+01 5.0555      4
## S2         2.0730e+01 1.6878e+01 5.0460      6
## FH_p=1_q=1 4.7555e+00 1.3376e+00 4.1119      3
```

```
## NULL
```

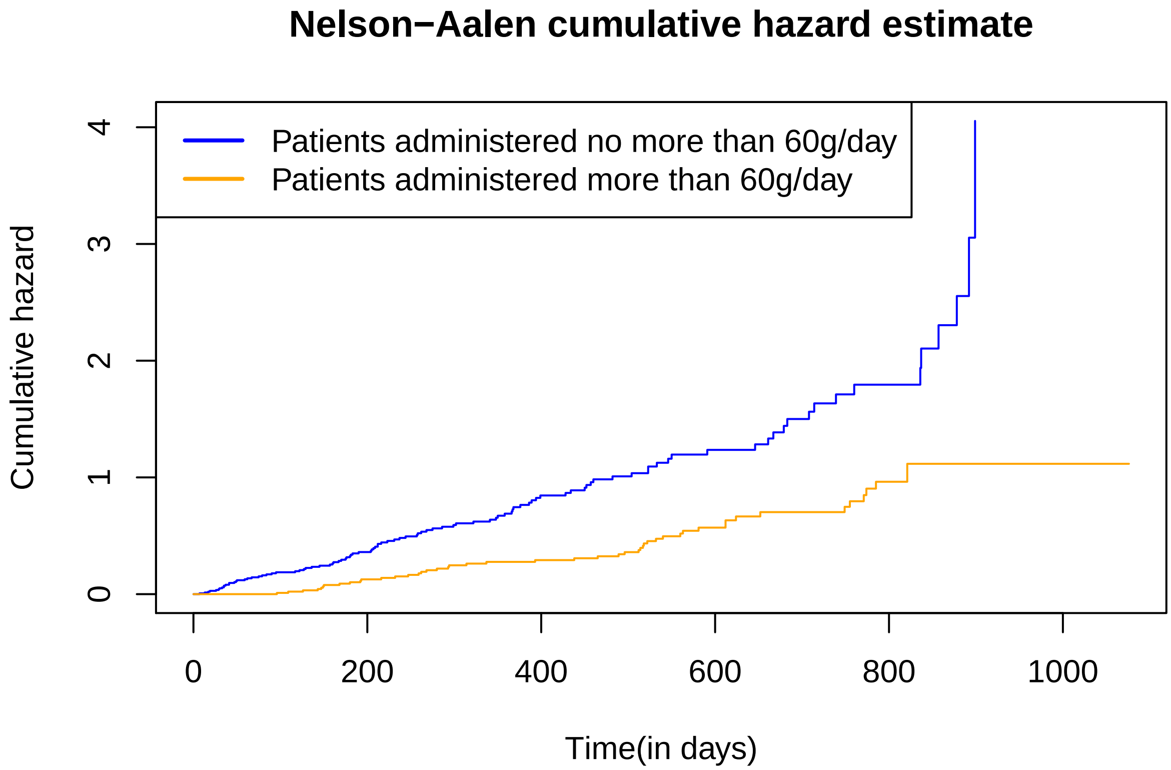
```
(1-pnorm(abs(-4.9534)))*2
```

```
## [1] 7.292784e-07
```

Based on the calculation, we can see that the Wilcoxon-Gehan-Breslow test gave us a Z-value of -4.9534 which gave us a two-sided p-value that is 7.293×10^{-7}. Therefore, based on the Wilcoxon-Gehan-Breslow test, we have evidence from the data to reject null hypothesis that there is no difference between the distribution of time until exit from maintenance in two groups with whether the participant have high dose of methadone.

Then we are going to plot the estimated hazard functions for patients with and without having more than 60 mg/day of methadone.

```
plot(km.addicts.dose, col = c("blue", "orange"), fun = "cumhaz",xlab = "Time(in days)",
     main = "Nelson-Aalen cumulative hazard estimate", ylab = "Cumulative hazard")
legend("topleft", c("Patients administered no more than 60g/day",
                    "Patients administered more than 60g/day"),
       col = c("blue","orange"), lwd = c(2, 2))
```



As we can see from the plot, the cumulative hazard distribution doesseem different and the hazard function are not crossing. Therefore, it may lead the log-rank test has similar p-value as Wilcoxon-Gehan-Breslow tests gave us based on the two test statistics magnitude. These two curves differ enough to for both two tests to detect the significantly difference.

Part(e)

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

```
## Call:
## survdiff(formula = s.addicts ~ prison + strata(clinic), data = addicts)
##
##            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
```

Our null hypothesis for stratified logrank test is the distribution of the time until exit from maintenance dose not differ by history of previous incarceration if the participants are from same clinic for all clinic.

Our null hypothesis for standard logrank test is the distribution of the time until exit from maintenance dose not differ by history of previous incarceration. The difference is the standard logrank test is not comparing the individuals with different incarceration history inside one clinic, which is not adjusting for the potential confounding effect of clinic.

Our alternative hypothesis for the stratified logrank test is at least for one clinic, the distribution of the time until exit from maintenance differ by history of precious incarceration even the participants are from same clinic.

Our alternative hypothesis for the standard logrank test is the distribution of the time until exit from maintenance differ by history of precious incarceration. The difference is the standard logrank test does not have the condition of comparing individuals within the same clinic.

Based on a stratified logrank test, we have a chi-square test statistic of 4 and associated p-value of 0.04. Therefore, the data does not provide enough evidence for us to reject the null hypothesis for stratified logrank test at the significance level of 0.05.

Part(f)

We estimated the median residual time by finding the time that is having half survival rate as the time already does and use that time to subtract with the time. The table is showing my estimation and the estimation calculated by the designed R function.

```
source("getmedianres.R")
calc_res = function(time){
  s = summary(km.addicts, times = time)$surv * 0.5
  med_time = min(km.addicts$time[km.addicts$surv < s])
  res_time = med_time - time
  return(res_time)
}
tab2 = matrix(nrow = 3, ncol =6)
tab2[,1] = c(4,8,12)
tab2[,2] = c(120,240,365)
tab2[1,3] = calc_res(120)
tab2[2,3] = calc_res(240)
tab2[3,3] = calc_res(365)
getrem.120 =getmedianres(s.addicts, times = 120, confint = TRUE)
getrem.240 = getmedianres(s.addicts, times = 240, confint = TRUE)
getrem.365 = getmedianres(s.addicts, times = 365, confint = TRUE)
tab2[1,4:6] = c(getrem.120$estimates, getrem.120$ci.lower, getrem.120$ci.upper)
tab2[2,4:6] = c(getrem.240$estimates, getrem.240$ci.lower, getrem.240$ci.upper)
tab2[3,4:6] = c(getrem.365$estimates, getrem.365$ci.lower, getrem.365$ci.upper)
colnames(tab2) = c("Months", "Days", "Kaplan-Meier Estimator","R Estimator","95% CI Low","95% CI High")
kable(tab2)
```

Months	Days	Kaplan-Meier Estimator	R Estimator	95% CI Low	95% CI High
4	120	420	420	376	532
8	240	427	427	323	520
12	365	384	384	296	456