Homework Assignment #4, Applied Survival Analysis
Due Monday February 21, 2022

1. Consider the following survival data for a two-group study comparing control to treatment:

Subject	Survival Time	Censoring Indicator	Group
1	5	1	C
2	8	1	C
3	7	1	T
4	2	1	T
5	8	0	T

In class we plotted these data as time lines indicating the failure time for each subject. We also expressed the data as a series of four 2-by-2 tables, one for each failure time, in the following form:

	control	treatment	
failure	d_{0i}	d_{1i}	d_i
Non-failure	$n_{0i} - d_{0i}$	$n_{1i} - d_{1i}$	$n_i - d_i$
	n_{0i}	n_{1i}	n_i

These results are on the class Canvas page in the class 4 module, with file name

“solution to exercise Applied Survival Analysis Survival as twobytwo tables.docx”

a. Find the Kaplan-Meier estimate of the survival function, *ignoring group*. Then plot the estimate.

t = 2			
	C	T	Total
Failure	0	1	1
Non-failure	2	2	4
Total	2	3	5

t = 5			
	C	T	Total
Failure	1	0	1
Non-failure	1	2	3
Total	2	2	4

t = 7			
	C	T	Total
Failure	0	1	1
Non-failure	1	1	2
Total	1	2	3

t = 8			
	C	T	Total
Failure	1	0	1
Non-failure	0	1	1
Total	1	1	2

t_i	n_i	d_i	N_{oi}	d_{oi}	n_{1i}	d_{1i}	e_{oi}	e_{1i}	v_{oi}
2	5	1	2	0	3	1	0.4	0.6	0.24
5	4	1	2	1	2	0	0.5	0.5	0.25
7	3	1	1	0	2	1	0.333	0.666	0.22
8	2	1	1	1	1	0	0.5	0.5	0.25
sum				2		2	1.7333	2.2667	0.962

```
library(survival)
tm <- c(5,8,7,2,8)
cens <- c(1,1,1,1,0)
grp <- c(0,0,1,1,1)
result.km <- survfit(Surv(tm, cens) ~ 1)
summary(result.km)
plot(result.km, main = "Kaplan Meier Survival Curve", xlab = "Time", ylab = "Survival
Probability", col= "red")

Call: survfit(formula = Surv(tm, cens) ~ 1)

 Time    n.risk.   n.event.    survival.   std.err.    lower 95% CI.    upper 95% CI
   2        5         1          0.8        0.179        0.5161              1
   5        4         1          0.6        0.219        0.2933              1
   7        3         1          0.4        0.219        0.1367              1
   8        2         1          0.2        0.179        0.0346              1
```

Compute the (1) log-rank, (2) Gehan, and (3) Prentice modified Gehan test statistics comparing C to T. For the later, you will need to first obtain $\hat{S}(t)$ (ignoring group indicator); you may do this in R if you like.

(a) The weighted log-rank test

```
library(survival)
tm <- c(5,8,7,2,8)
cens <- c(1,1,1,1,0)
grp <- c(0,0,1,1,1)
survdiff(Surv(tm, cens) ~ grp, rho = 0)
Call:
```

```
survdiff(formula = Surv(tm, cens) ~ grp)

              N      Observed.    Expected        (O-E)^2/E      (O-E)^2/V
grp=0         2          2           1.73           0.0410          0.0739
grp=1         3          2           2.27            0.0314         0.0739

 Chisq= 0.1 on 1 degrees of freedom, p= 0.8
```

t_i	n_i	d_i	N_{oi}	d_{oi}	n_{1i}	d_{1i}	e_{oi}	e_{1i}	v_{oi}
2	5	1	2	0	3	1	0.4	0.6	0.24
5	4	1	2	1	2	0	0.5	0.5	0.25
7	3	1	1	0	2	1	0.333	0.666	0.22
8	2	1	1	1	1	0	0.5	0.5	0.25
sum				2		2	1.7333	2.2667	0.962

The P value is 0.8, indicating that the group difference is not statistically significant (which is not surprising due to the extremely small sample size in this example)

(b) Gehan test statistic

t_i	n_i	d_i	N_{oi}	d_{oi}	n_{1i}	d_{1i}	e_{oi}	e_{1i}	v_{oi}	u_o
2	5	1	2	0	3	1	0.4	0.6	0.24	-2
5	4	1	2	1	2	0	0.5	0.5	0.25	2
7	3	1	1	0	2	1	0.333	0.666	0.22	-1
8	2	1	1	1	1	0	0.5	0.5	0.25	1
sum				2		2	1.7333	2.2667	0.962	

$$\chi^2 \ = \ 0 \quad df \ = \ 1 \; and \; p - value \ = 1$$

The P value is 1, indicating that the group difference is not statistically significant (which is not surprising due to the extremely small sample size in this example)

(c) Prentice modified Gehan test statistic

```
library(survival)
tm <- c(5,8,7,2,8)
cens <- c(1,1,1,1,0)
grp <- c(0,0,1,1,1)
survdiff(Surv(tm, cens) ~ grp, rho = 1)

survdiff(formula = Surv(tm, cens) ~ grp, rho = 1)

              N        Observed      Expected         (O-E)^2/E           (O-E)^2/V
grp=0         2           1.2           1.2               0                   0
grp=1         3            1.6          1.6               0                   0

 Chisq= 0  on 1 degrees of freedom, p= 1
```

t_i	n_i	d_i	N_{oi}	d_{oi}	n_{1i}	d_{1i}	e_{oi}	e_{1i}	v_{oi}	u_o
2	5	1	2	0	3	1	0.4	0.6	0.24	-2
5	4	1	2	1	2	0	0.5	0.5	0.25	2
7	3	1	1	0	2	1	0.333	0.666	0.22	-1
8	2	1	1	1	1	0	0.5	0.5	0.25	1
sum				2		2	1.7333	2.2667	0.962	

$$\chi^2 \; = \; 0 \quad df \; = \; 1 \; and \; p - value \; = 1$$

b. Carry out the log-rank test in R, as follows, and compare to your answer in part c. Also fit the Prentice modified Gehan test (with the option rho=1), and compare to part c above. Refer to the supplemental slides “supplement Class 4 Urn hypergeometric …”, on Canvas, Class 4 module, which gives formulas for the mean and variance of d_{0i}.

```
tm <- c(5, 8, 7, 2, 8)
```

```
cens <- c(1, 1, 1, 1, 0)
grp <- c(0,0, 1, 1, 1)

library(survival)
result <- survdiff(Surv(tm, cens) ~ grp, rho=0)
summary(result)

Call:
survdiff(formula = Surv(tm, cens) ~ grp, rho = 0)

          N     Observed    Expected    (O-E)^2/E    (O-E)^2/V
grp=0     2         2          1.73        0.0410       0.0739
grp=1     3         2          2.27        0.0314       0.0739

 Chisq= 0.1 on 1 degrees of freedom, p= 0.8
```

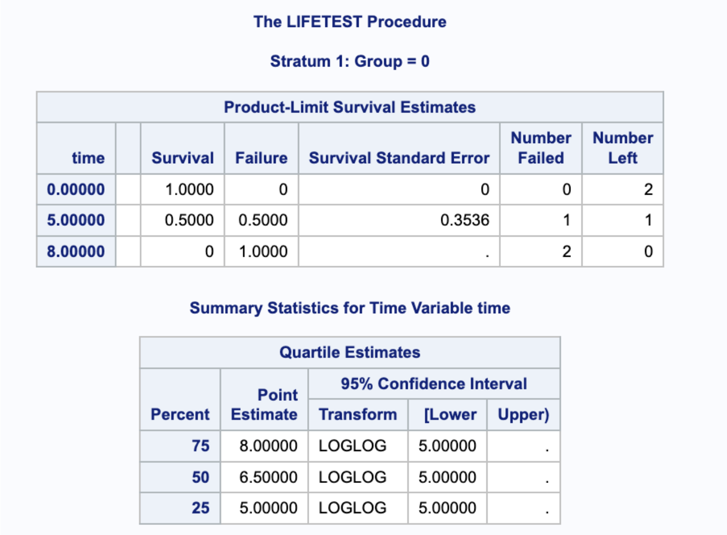
Fitting the Prentice modified Gehan test (with the option rho=1),

```
tm <- c(5, 8, 7, 2, 8)
cens <- c(1, 1, 1, 1, 0)
grp <- c(0,0, 1, 1, 1)
library(survival)
survdiff(Surv(tm, cens) ~ grp, rho=1)
Call:
survdiff(formula = Surv(tm, cens) ~ grp, rho = 1)

          N      Observed       Expected     (O-E)^2/E      (O-E)^2/V
grp=0     2        1.2            1.2            0              0
grp=1     3        1.6            1.6            0              0

 Chisq= 0  on 1 degrees of freedom, p= 1
```

It appears that the p-value and chi-square value for this and what is in part (c) are the same.

f. Fit this model in SAS. Show that you get the same answer.

```
proc lifetest data=example plots=(s) graphics;
  time time*cens(0);
  strata grp \ TEST=(LOGRANK WILCOXON FLEMING(1,0)) ;
run;
```



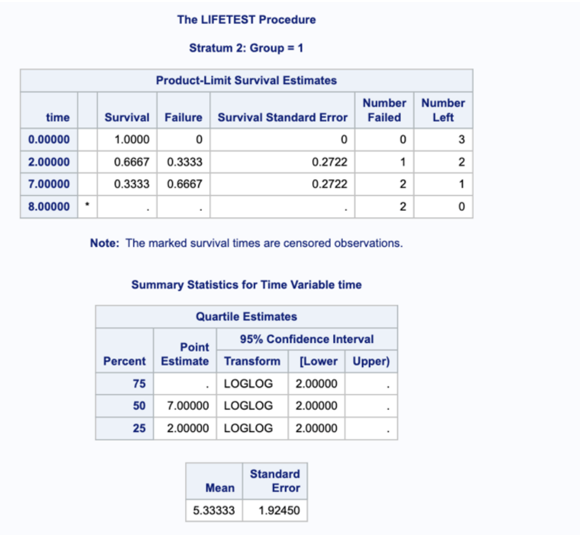
The LIFETEST Procedure

Stratum 1: Group = 0

Product-Limit Survival Estimates						
time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00000		1.0000	0	0	0	2
5.00000		0.5000	0.5000	0.3536	1	1
8.00000		0	1.0000	.	2	0

Summary Statistics for Time Variable time

Quartile Estimates					
		95% Confidence Interval			
Percent	Point Estimate	Transform	[Lower	Upper)
75	8.00000	LOGLOG	5.00000	.
50	6.50000	LOGLOG	5.00000	.
25	5.00000	LOGLOG	5.00000	.



The LIFETEST Procedure

Stratum 2: Group = 1

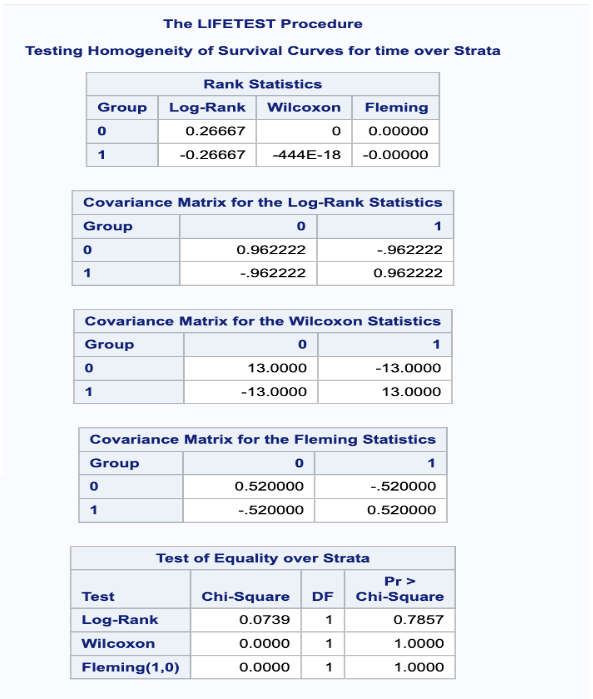
Product-Limit Survival Estimates						
time		Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00000		1.0000	0	0	0	3
2.00000		0.6667	0.3333	0.2722	1	2
7.00000		0.3333	0.6667	0.2722	2	1
8.00000	*	.	.	.	2	0

Note: The marked survival times are censored observations.

Summary Statistics for Time Variable time

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	.	LOGLOG	2.00000	.
50	7.00000	LOGLOG	2.00000	.
25	2.00000	LOGLOG	2.00000	.

Mean	Standard Error
5.33333	1.92450



The LIFETEST Procedure

Testing Homogeneity of Survival Curves for time over Strata

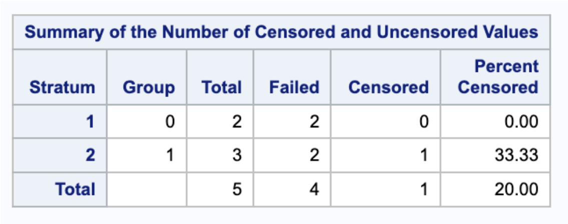
Rank Statistics			
Group	Log-Rank	Wilcoxon	Fleming
0	0.26667	0	0.00000
1	-0.26667	-444E-18	-0.00000

Covariance Matrix for the Log-Rank Statistics		
Group	0	1
0	0.962222	-.962222
1	-.962222	0.962222

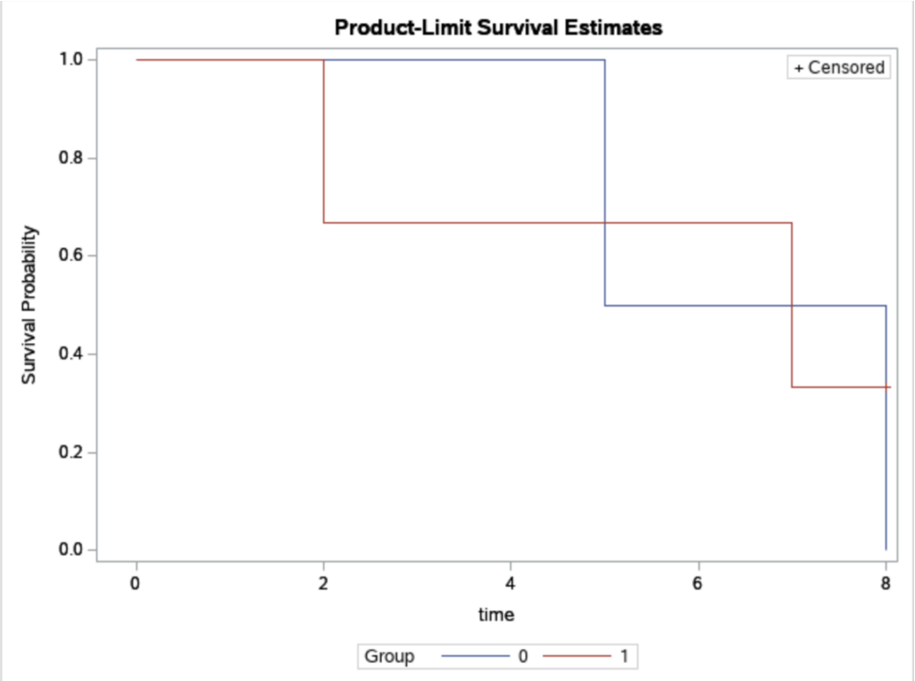
Covariance Matrix for the Wilcoxon Statistics		
Group	0	1
0	13.0000	-13.0000
1	-13.0000	13.0000

Covariance Matrix for the Fleming Statistics		
Group	0	1
0	0.520000	-.520000
1	-.520000	0.520000

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.0739	1	0.7857
Wilcoxon	0.0000	1	1.0000
Fleming(1,0)	0.0000	1	1.0000



Summary of the Number of Censored and Uncensored Values					
Stratum	Group	Total	Failed	Censored	Percent Censored
1	0	2	2	0	0.00
2	1	3	2	1	33.33
Total		5	4	1	20.00



The answer is also same here.

2. a. Using the “pancreatic2” data from the “asaur” package, use R to compare the PFS for the locally advanced and metastatic groups using the log-rank test (rho=0) and then the Prentice modification of the Gehan test (rho=1). Explain the difference in the p-values.

Using the log-ranked test, we have

```
library(asaur)
attach(pancreatic2)
survdiff(Surv(pfs) ~ stage, rho = 0)

Call:
survdiff(formula = Surv(pfs) ~ stage, rho = 0)

             N     Observed     Expected     (O-E)^2/E     (O-E)^2/V
stage= LA    8        8           12.3         1.49           2.25
stage=M     33       33            28.7        0.64            2.25

 Chisq= 2.2 on 1 degrees of freedom, p= 0.1
```

It can be realized that the number of patients in each group equals the corresponding observed number of events, since there is no censoring. The value of the chi squared *statistic is 2.2 with 1 degree of freedom, and the P value is 0.1 which is not statically significant.*

Using the Prentice modification of the Gehan test, we have

```
library(asaur)
attach(pancreatic2)
survdiff(Surv(pfs) ~ stage, rho = 1)

Call:
survdiff(formula = Surv(pfs) ~ stage, rho = 1)

           N Observed Expected (O-E)^2/E (O-E)^2/V
stage=LA  8     2.34     5.88     2.128      4.71
stage=M  33    18.76    15.22     0.822      4.71

 Chisq= 4.7  on 1 degrees of freedom, p= 0.03
```

We obtained a p- value of 0.03 what is statically significant at the 5% level. It appears that the Prentice modification places higher weight on earlier survival times.

c. Since there is no censoring in this data set, we may use a Wilcoxon rank test to compare the two groups. In R, use the function "wilcox.test" to do this comparison. Which of the two tests in part a is the Wilcoxon test most similar too? Explain.

```
library(asaur)

attach(pancreatic2)

wilcox.test(pancreatic2$pfs ~ pancreatic2$stage)

Wilcoxon rank sum test with continuity correction

data:  pancreatic2$pfs by pancreatic2$stage

W = 204.5, p-value = 0.01783

alternative hypothesis: true location shift is not equal to 0
```

Explanation. By using the Wilcoxon rank test to compare the two groups it appears that their work oxen test is similar to the Prentice test. This is because move test appears to be statistically significant

d. One could in principle carry out a two-sample t-test to compare these two survival curves since there is no censoring. Explain why this would not be a good approach.

This is not a good approach because it will yield a biased survival distribution and an incorrect P- value.

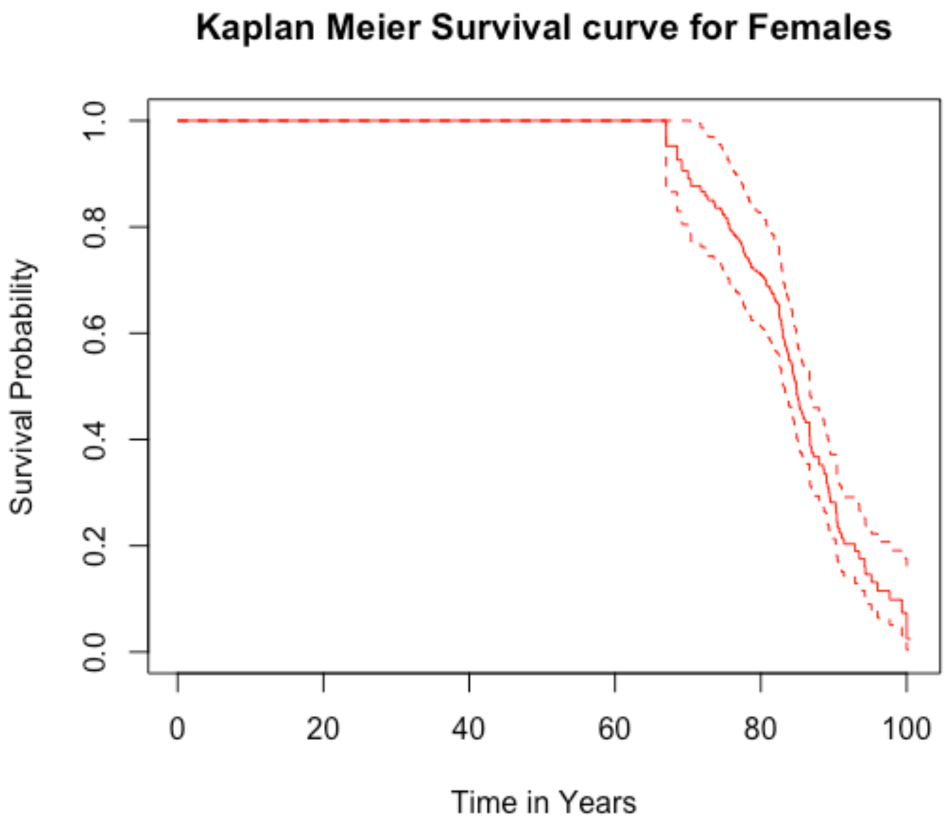
3. Fit the model from class to the “ChanningHouse” data (in the “asaur” package), accounting for left truncation. Plot the Kaplan-Meier and Nelson-Aalen survival curve estimates separately for men and women. Here is a start:

```
result <- survfit(Surv(entry, exit, cens) ~ 1, data=ChanningHouse,
    subset={sex=="Female"})
```

```
library(asaur)
library(survival)

yearentry <- ChanningHouse$entry/12
yearexit <- ChanningHouse$exit/12
yeartime <- ChanningHouse$time/12
```

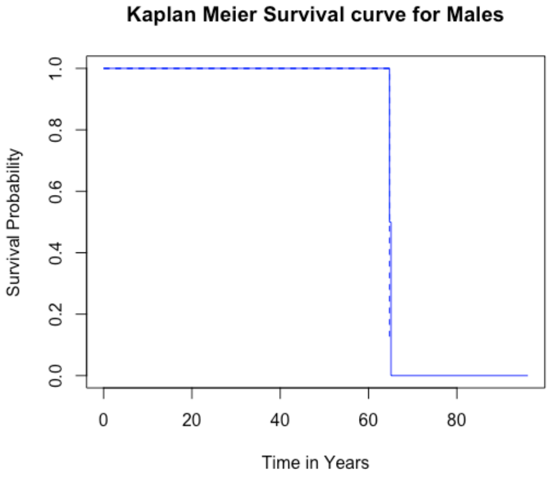
Kaplan-Meier Survival Curve for Females

```
result <- survfit(Surv(yearentry, yearexit, cens)~1,type= "kaplan-meier",  data = ChanningHouse,
              subset={sex=="Female"})
plot(result, main= "Kaplan Meier Survival curve for Females", xlab = "Time in Years",
    ylab = "Survival Probability", col = "red")
```



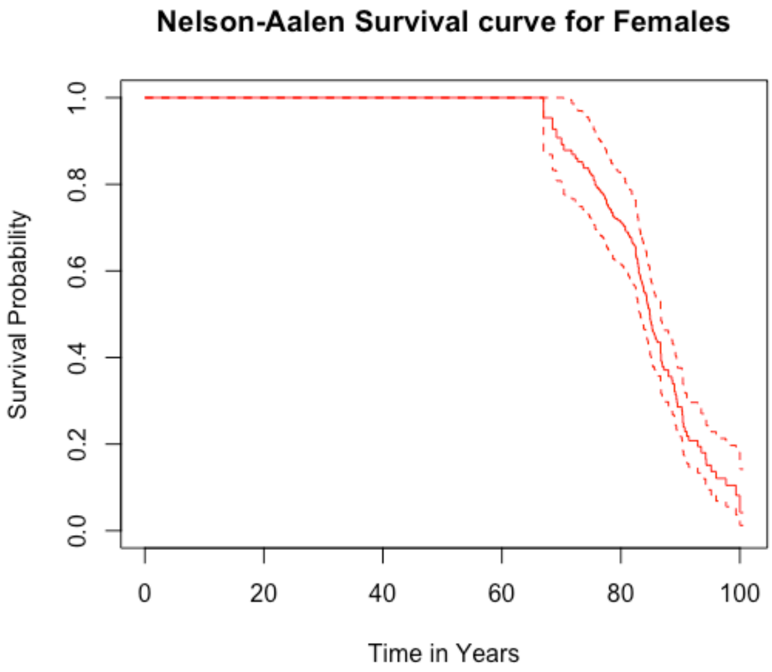
Kaplan-Meier Survival Curve for Males

```
result <- survfit(Surv(yearentry, yearexit, cens)~1,type= "kaplan-meier",  data = Channi
ngHouse,
                subset={sex=="Male"})
plot(result, main= "Kaplan Meier Survival curve for Males", xlab = "Time in Years",
    ylab = "Survival Probability", col = "blue")
```



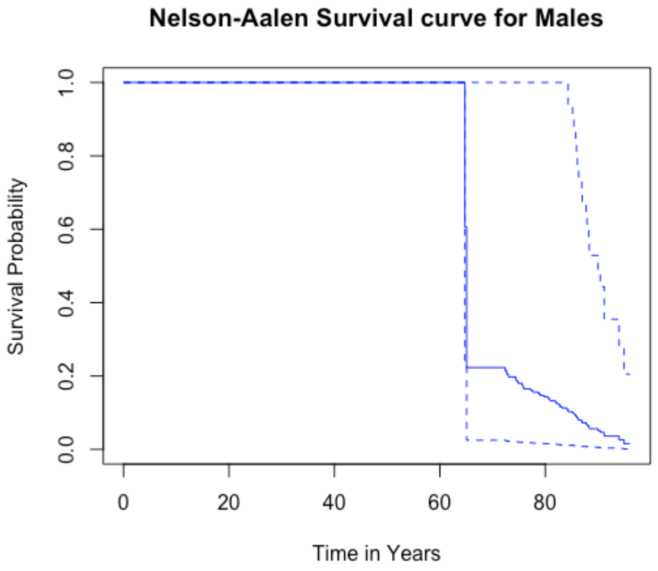
Nelson-Aalen Survival Curve for women

```
result <- survfit(Surv(yearentry, yearexit, cens)~1,type = "fleming-harrington",  data =
ChanningHouse,
                subset={sex=="Female"})
plot(result, main= "Nelson-Aalen Survival curve for Females", xlab = "Time in Years",
    ylab = "Survival Probability", col = "red")
```



Nelson-Aalen Survival Curve for males

```
result <- survfit(Surv(yearentry, yearexit, cens)~1,type = "fleming-harrington",  data =
ChanningHouse,
                subset={sex=="Male"})
plot(result, main= "Nelson-Aalen Survival curve for Males", xlab = "Time in Years",
    ylab = "Survival Probability", col = "blue")
```



BUT, please convert time (which is in months) into years before you fit the model and plot the results.

Next, show how you can get a more informative plot for the men. (Hint: condition on surviving a certain amount of time by removing shorter times.)

4. Consider the "ovarian" data set that is included in the R "survival" package. The survival times are "ovarian$futime" and the censoring indicators are "ovarian$fustat".
   a. Find the Kaplan-Meier survival curve and plot it. Ignore the other variables in the data set.

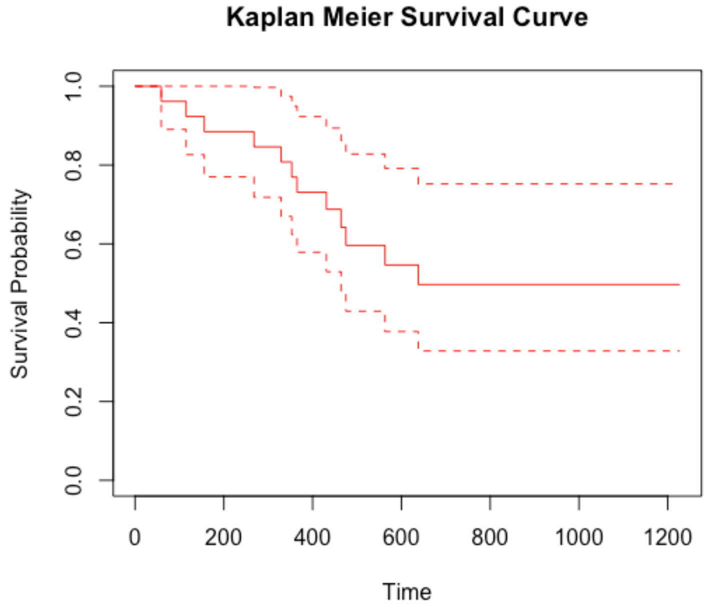
library(asaur)

library(survival)

result.km <- survfit(Surv(futime, fustat) ~ 1, data = ovarian)

summary(result.km)

plot(result.km, main = "Kaplan Meier Survival Curve", xlab = "Time",ylab = "Survival Probability", col= "red")



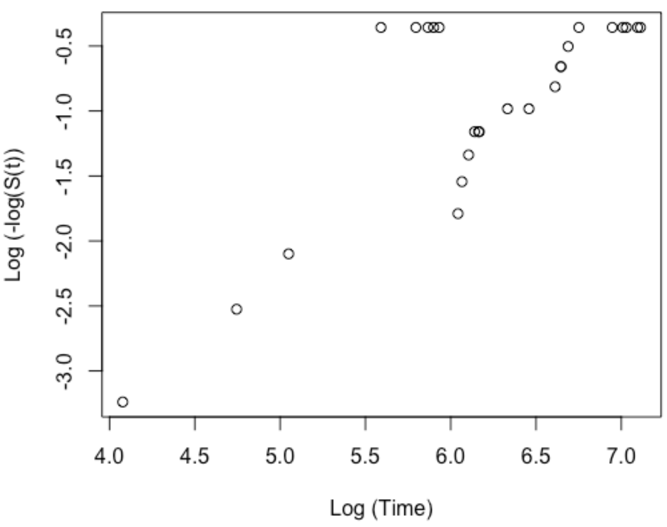
b. Does this survival curve follow a Weibull distribution? Check using the formula on slide 15 from the last lecture.

In R, we compute :

result.km_ovarian <- result.km$surv

plot(log(ovarian$futime), log(-log(result.km_ovarian)), xlab ="Log (Time)", ylab = "Log (-log(S(t))")

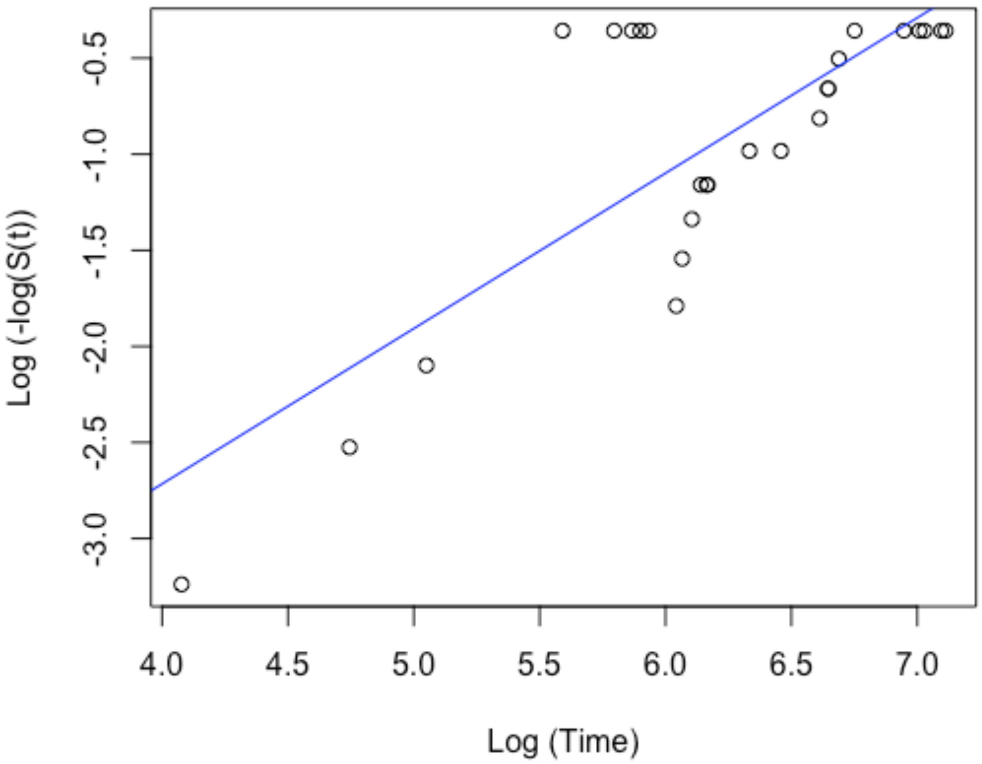
abline(lm(log(-log(result.km_ovarian))~ log(ovarian$futime)), col= "blue")



This shows that the survival curve does not follow a Weibull distribution. If if it were to follow a Weibull distribution, The data points would approximately follow the straight line.

c. Fit a straight line through the data, as you have done before, using the “lm” function.

ovarian_model <-lm(log(-log(result.km_ovarian))~ log(ovarian$futime))
ovarian_model



d. Report the values of α and λ based on the linear fit in part c.

Call:

```
lm(formula = log(-log(result.km_ovarian)) ~ log(ovarian$futime)
```

Coefficients:

```
(Intercept)  log(ovarian$futime)
   -5.9510               0.8088
```

$y = -5.9510 + \quad 0.8088$

$Intercept = ln(\lambda) = -5.9510\ \ and\ therefore\ \lambda = \ 0.00260$

$Slope\ \alpha = \ 0.8088$

e. Does the survival data follow an exponential distribution? How can you tell from the linear model fit?

The survival data do not follow an exponential distribution, because the slope which is constant does to accurately capture the general trend of data points. if the slope indicated by the points were closely matching the line of best fit, then we could say the survival data follows an exponential distribution.