Northeastern
College of Science

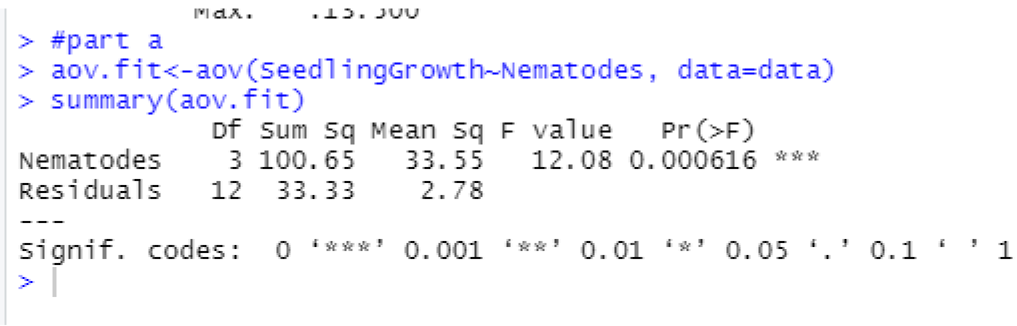
Applied Statistics
FINAL HOMEWORK

Anjali Patil
NUID: -00109806
4/28/2021

QUESTION 4:-

Part a:-

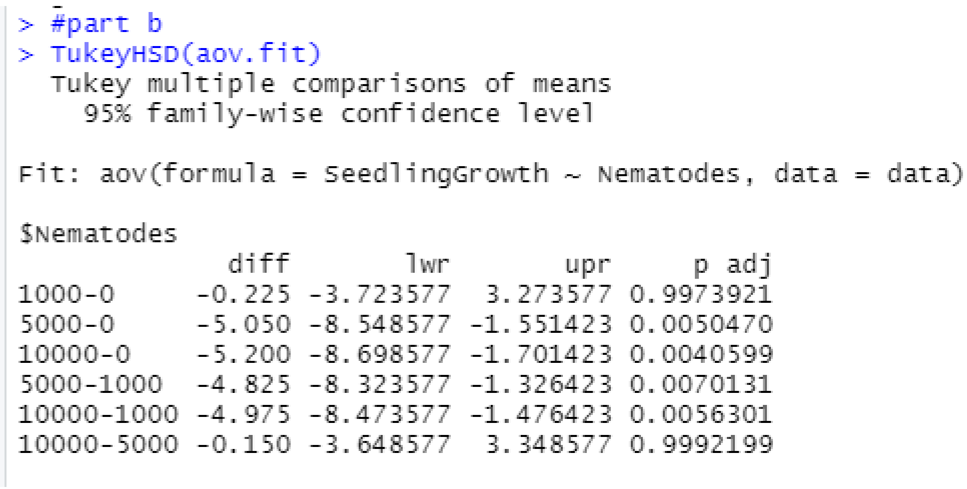
Below is code and output:-



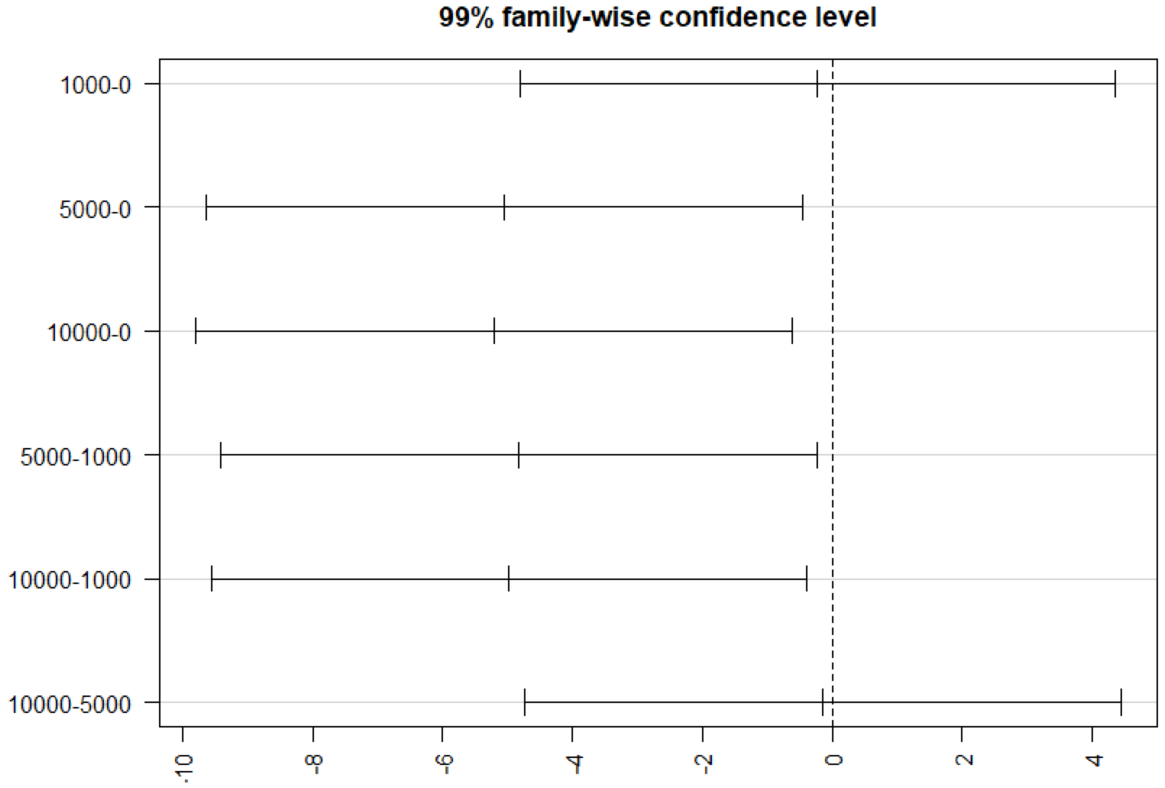
```
              Max.   :13.500
> #part a
> aov.fit<-aov(SeedlingGrowth~Nematodes, data=data)
> summary(aov.fit)
            Df Sum Sq Mean Sq F value   Pr(>F)    
Nematodes    3 100.65   33.55   12.08 0.000616 ***
Residuals   12  33.33    2.78                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Since p-value is 0.0006< 0.05, we conclude that nematodes have a significant effect on the plant growth at 0.05 level of significant.

Part b :-

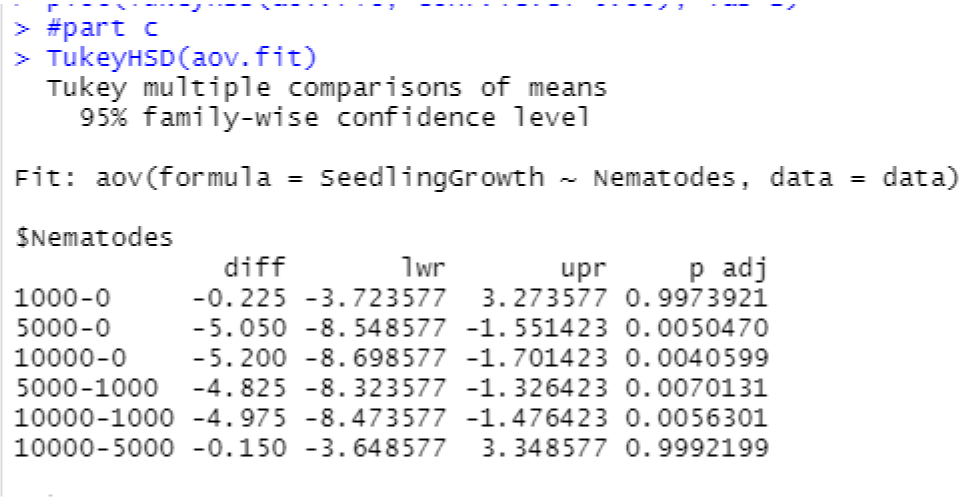


```
> #part b
> TukeyHSD(aov.fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SeedlingGrowth ~ Nematodes, data = data)

$Nematodes
             diff       lwr       upr     p adj
1000-0     -0.225 -3.723577  3.273577 0.9973921
5000-0     -5.050 -8.548577 -1.551423 0.0050470
10000-0    -5.200 -8.698577 -1.701423 0.0040599
5000-1000  -4.825 -8.323577 -1.326423 0.0070131
10000-1000 -4.975 -8.473577 -1.476423 0.0056301
10000-5000 -0.150 -3.648577  3.348577 0.9992199
```



So here, since 5000-0, 10000-0, 5000-1000, 10000-1000 all have p-values less than 0.05, we declare those to be significant at 0.05 level of significant. Our data suggest that the 5000 and 10,000 nematode treatments both reduce seedling growth vs. the 0 nematode treatment, and they both reduce seedling growth vs. the 1,000 nematode treatment. Thus we conclude that nematodes reduce planet growth.

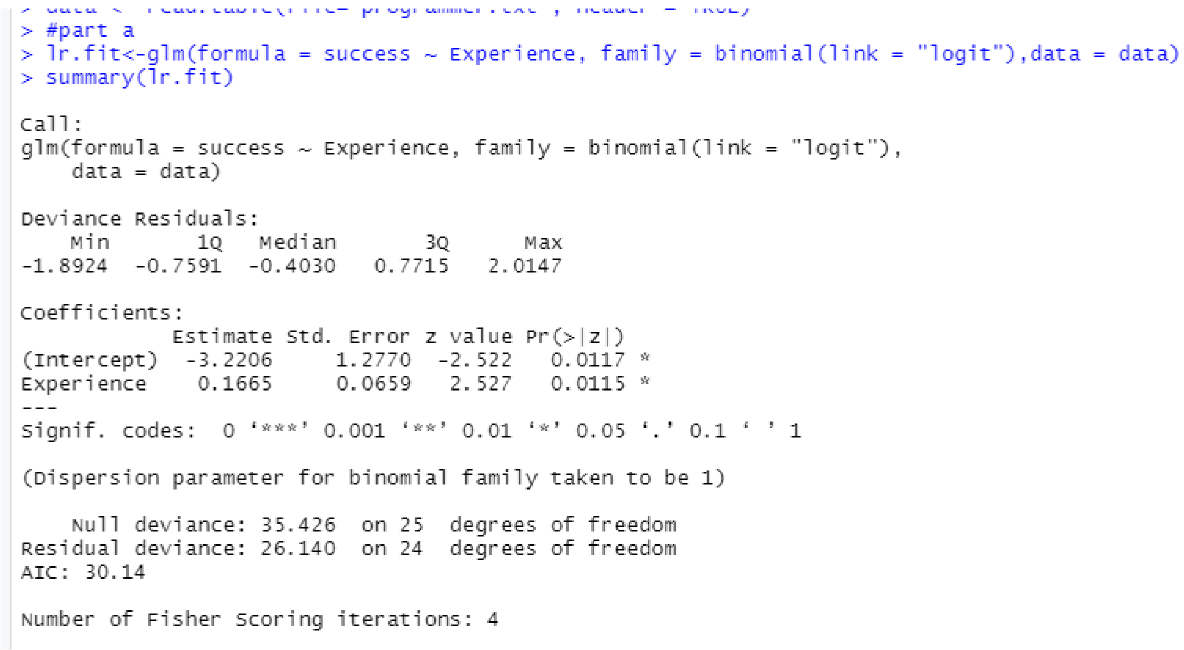
Part c : -



```
> #part c
> TukeyHSD(aov.fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SeedlingGrowth ~ Nematodes, data = data)

$Nematodes
             diff       lwr       upr     p adj
1000-0     -0.225 -3.723577  3.273577 0.9973921
5000-0     -5.050 -8.548577 -1.551423 0.0050470
10000-0    -5.200 -8.698577 -1.701423 0.0040599
5000-1000  -4.825 -8.323577 -1.326423 0.0070131
10000-1000 -4.975 -8.473577 -1.476423 0.0056301
10000-5000 -0.150 -3.648577  3.348577 0.9992199
```

QUESTION 5:-

Part a: -
We will use Logistic regression to carry out the statistical analysis
H0=There is no association between work experience and programmer’s ability
HA=There is association between work experience and programmer’s ability

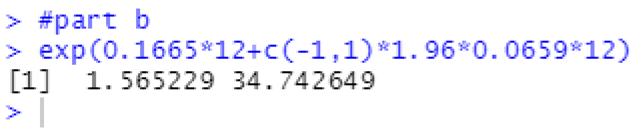


```
> data <- read.table(file= "programmer.txt", header = TRUE)
> #part a
> lr.fit<-glm(formula = success ~ Experience, family = binomial(link = "logit"),data = data)
> summary(lr.fit)

Call:
glm(formula = success ~ Experience, family = binomial(link = "logit"), 
    data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8924  -0.7591  -0.4030   0.7715   2.0147  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  -3.2206     1.2770  -2.522   0.0117 *
Experience    0.1665     0.0659   2.527   0.0115 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.426  on 25  degrees of freedom
Residual deviance: 26.140  on 24  degrees of freedom
AIC: 30.14

Number of Fisher Scoring iterations: 4
```

$\ln\left(\frac{\hat{P}}{1-\hat{P}}\right)$ =**-3.2206 +0.1665x1**

The coefficient of experience is positive, indicating that the log odds of improving the ability of a programmer- and thus the probability p itself- is higher for individuals who have more experience relative to those with less . Further, it implies that with more Experience the log odds of improving programs ability increases by 0.1665. Also when log odds increases the probability p increases.
Further since the p-value is 0.011<0.05, we reject the null hypotheses at 0.05 level of significant and conclude that there is significant association between work experience and programs ability.

Part b:-



```
> #part b
> exp(0.1665*12+c(-1,1)*1.96*0.0659*12)
[1]  1.565229 34.742649
> 
```

Note we multiplied by 12 to change the unit to year

We are 95% confidence that , the interval in improvement in the odds of completing the task with specified time period for each extra year of work experience is (1.565 , 34.7426)

Part c:-

$\ln\left(\frac{\hat{P}}{1-\hat{P}}\right)$ =**-3.2206 +0.1665(24)= .7754**

$\left(\frac{\hat{P}}{1-\hat{P}}\right)$**=2.17**

p = 68.4% the probability of finishing the task within the period of time for an employee with 24 months of previous work experience. So this programmer should have a salary of $60,000 per year.

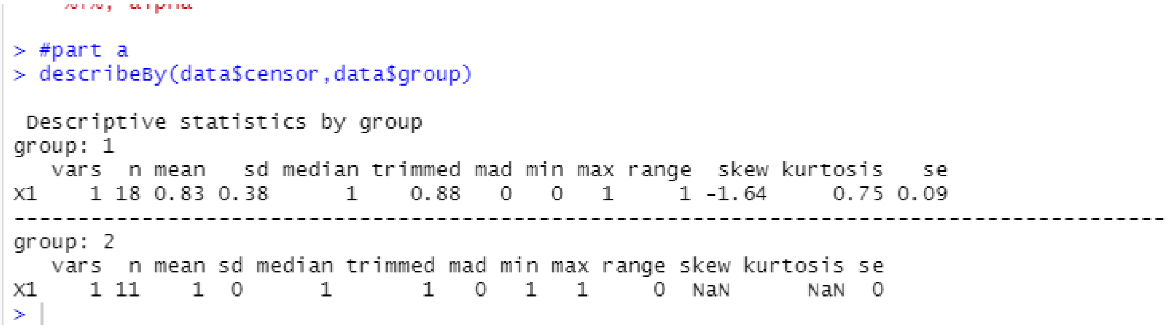
$\ln\left(\frac{\hat{P}}{1-\hat{P}}\right) =$ **-3.2206 +0.1665(18)= -0.223**

$\left(\frac{\hat{P}}{1-\hat{P}}\right)$**=0.8001**

p = 44.4% the probability of finishing the task within the period of time for an employee with 18 months of previous work experience From the analysis second programmer has 44.4% probability of finishing the task which is low compare to the employee with 24 month of work experience with 68.4% of finishing the task. The second programmer would be a better deal if company wants to save money but he/she will not be better deal to finish the task within the time period.

QUESTION 6:-

Part a: -

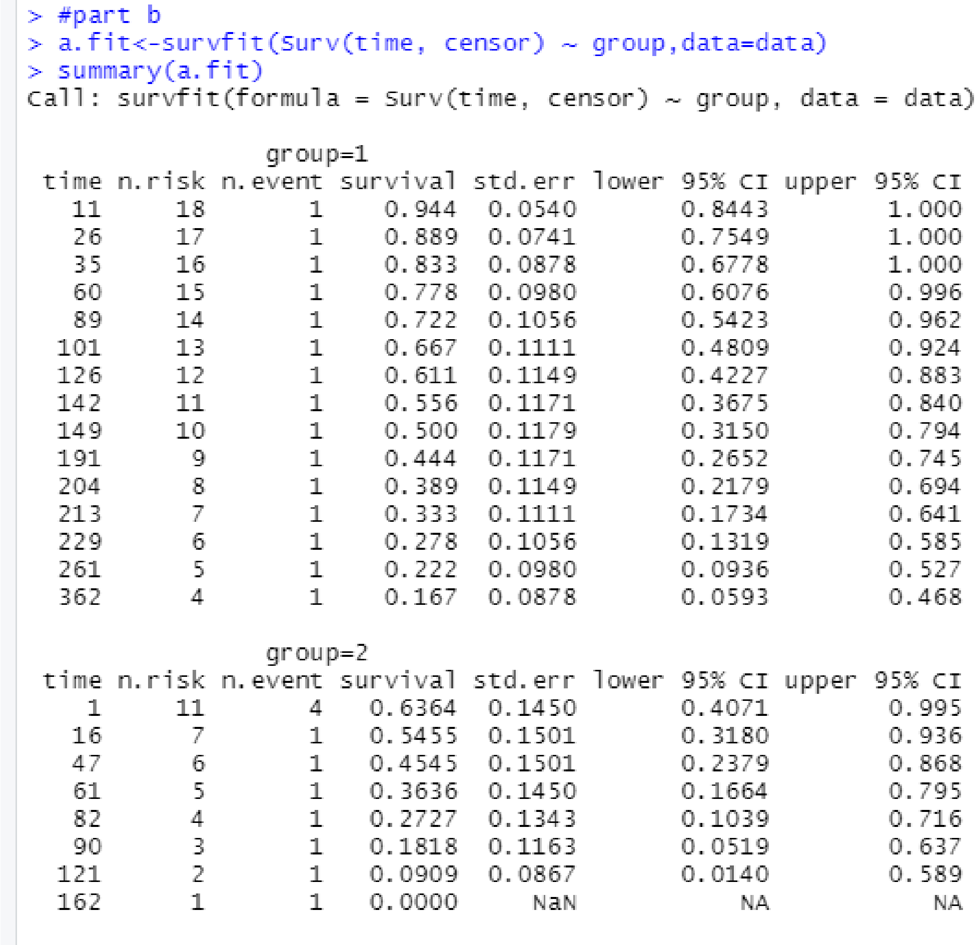


```
    %+%, alpha

> #part a
> describeBy(data$censor,data$group)

 Descriptive statistics by group 
group: 1
   vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
X1    1 18 0.83 0.38      1    0.88   0   0   1     1 -1.64     0.75 0.09
--------------------------------------------------------------------------------- 
group: 2
   vars  n mean sd median trimmed mad min max range skew kurtosis se
X1    1 11    1  0      1       1   0   1   1     0  NaN      NaN  0
> 
```

Group 1: 18 death occurred

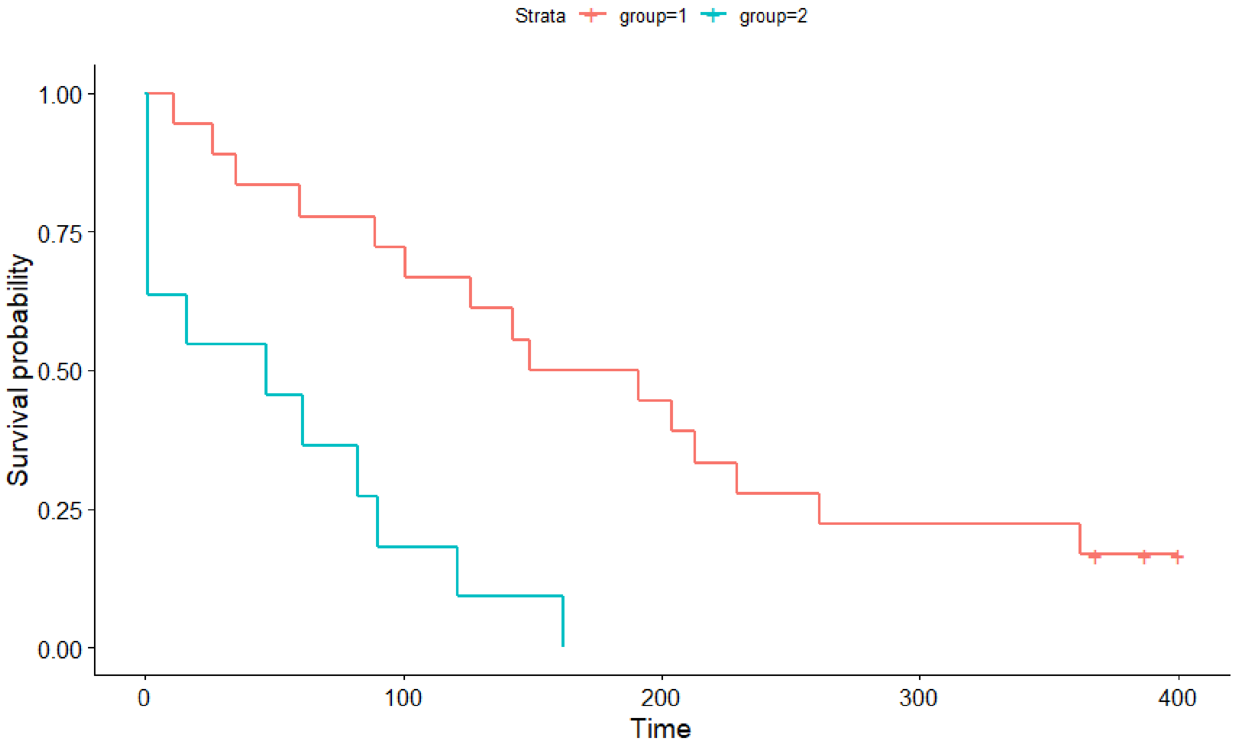
Group 2: 11 death occurred

Part b: -



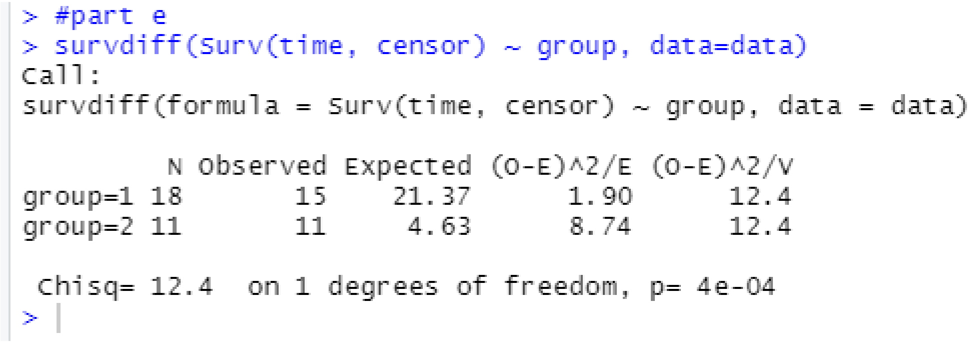
```
> #part b
> a.fit<-survfit(Surv(time, censor) ~ group,data=data)
> summary(a.fit)
Call: survfit(formula = Surv(time, censor) ~ group, data = data)

                group=1 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   11     18       1    0.944  0.0540       0.8443        1.000
   26     17       1    0.889  0.0741       0.7549        1.000
   35     16       1    0.833  0.0878       0.6778        1.000
   60     15       1    0.778  0.0980       0.6076        0.996
   89     14       1    0.722  0.1056       0.5423        0.962
  101     13       1    0.667  0.1111       0.4809        0.924
  126     12       1    0.611  0.1149       0.4227        0.883
  142     11       1    0.556  0.1171       0.3675        0.840
  149     10       1    0.500  0.1179       0.3150        0.794
  191      9       1    0.444  0.1171       0.2652        0.745
  204      8       1    0.389  0.1149       0.2179        0.694
  213      7       1    0.333  0.1111       0.1734        0.641
  229      6       1    0.278  0.1056       0.1319        0.585
  261      5       1    0.222  0.0980       0.0936        0.527
  362      4       1    0.167  0.0878       0.0593        0.468

                group=2 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     11       4   0.6364  0.1450       0.4071        0.995
   16      7       1   0.5455  0.1501       0.3180        0.936
   47      6       1   0.4545  0.1501       0.2379        0.868
   61      5       1   0.3636  0.1450       0.1664        0.795
   82      4       1   0.2727  0.1343       0.1039        0.716
   90      3       1   0.1818  0.1163       0.0519        0.637
  121      2       1   0.0909  0.0867       0.0140        0.589
  162      1       1   0.0000     NaN           NA           NA
```

Part c: -



Part d: -
Based on the curves, it appears that group 1 (patients treated with drug) survived longer than group 2 (patients with no drug treatment).

Part e: -



```
> #part e
> survdiff(Surv(time, censor) ~ group, data=data)
Call:
survdiff(formula = Surv(time, censor) ~ group, data = data)

         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 18       15    21.37      1.90      12.4
group=2 11       11     4.63      8.74      12.4

 Chisq= 12.4  on 1 degrees of freedom, p= 4e-04
> 
```

Since p-value is .0004 < 0.05, we reject null hypothesis at 0.05 level of significant. We conclude that the distributions of survival times are not identical in the two groups.