Assignment 3

Zhiwen Tan

3/14/2022

Question 1

a) To get the 95% confidence interval for median survival time, that is find $t_{0.5}$. We first need to fit a Weibull model for the data at stress level $750N/mm^2$, the result below is the model fit result. Then we have $\hat{\sigma}$, $\hat{\mu}$ and estimate $\hat{Var}$.Then let $Y = logT$, and we can calculate $y_{0.5}$. After this step, we can calculate the $Var(\hat{y_{0.5}})$. Now, we can use 95% CI for $y_{0.5}$. Finally, we can use $t_{0.5} = exp(y_{0.5})$ to get 95% confidence interval for median(t_0.5) is [3109.220, 8525.353].

```
##
## Call:
## survreg(formula = Surv(x, delta) ~ 1)
##              Value Std. Error     z      p
## (Intercept)  8.798      0.234 37.55 <2e-16
## Log(scale)  -0.376      0.267 -1.41   0.16
##
## Scale= 0.686
##
## Weibull distribution
## Loglik(model)= -86.8   Loglik(intercept only)= -86.8
## Number of Newton-Raphson Iterations: 6
## n= 10
```

```r
# a)
y <- log(-log(0.5))
y_mid <- fit.weib$coef + y*fit.weib$scale
vary_mid <- as.numeric(t(matrix(c(1, y*fit.weib$scale))) %*%
                         as.matrix(fit.weib$var) %*%
                         matrix(c(1, y*fit.weib$scale)))
CIL <- exp(y_mid - 1.96*sqrt(vary_mid))
CIU <- exp(y_mid + 1.96*sqrt(vary_mid))
CI <- c(CIL, CIU)
CI
```

```
## (Intercept) (Intercept)
##    3109.220    8525.353
```

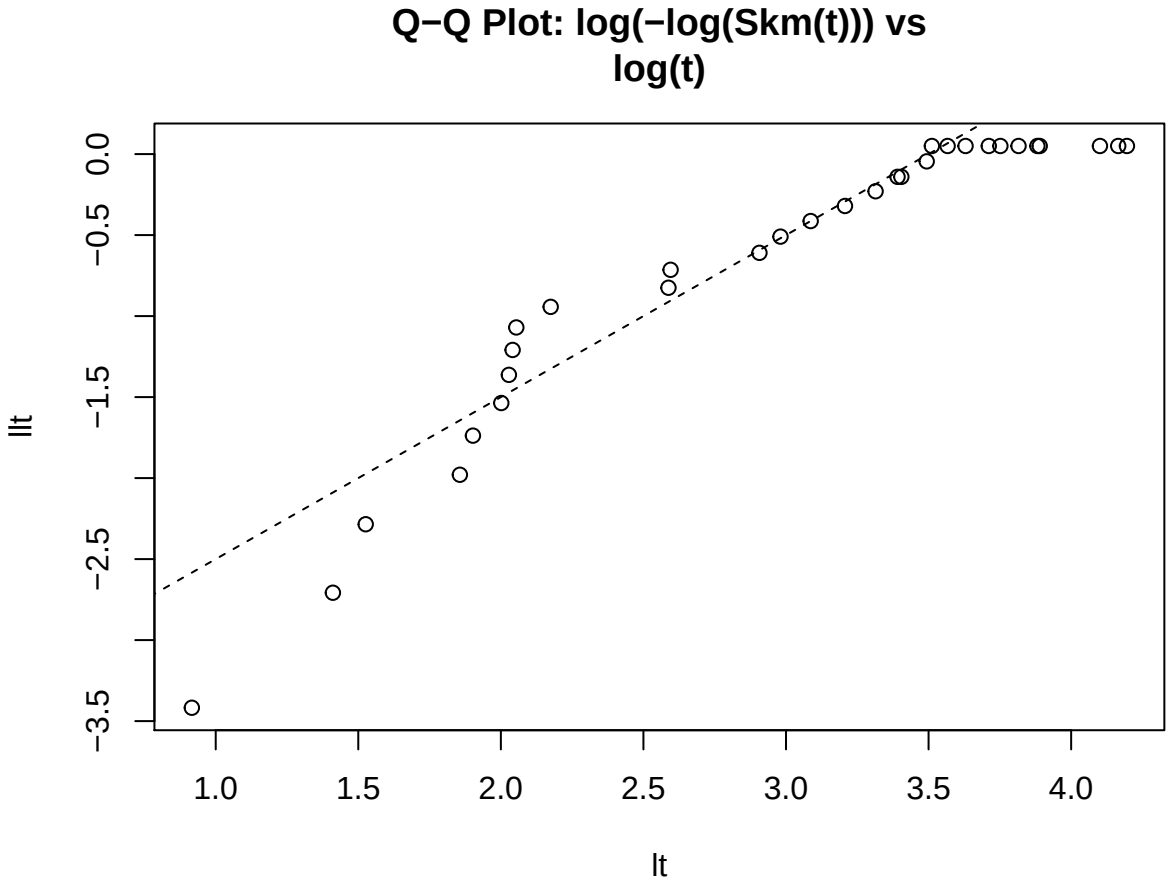
b) To get the confidence interval for the probability that the type of spring survives over 6000 thousand cycles at the stress level $750N = mm^2$. that is find the $S(6000)$. The first step is to find the CI for $log[-log(S(t))]$, this is equal to $(log(t)-\mu)e^{-\phi}$. Then we can find $var((log(t)-\mu)e^{-\phi})$ using δ-method. Now, we have the 95% CI for $log[-log(S(6000))]$, Then we can calculate CIl and CIu by using $exp(-exp(\hat{\psi}_u))$ and $exp(-exp(\hat{\psi}_l))$. Finally, we can get the result 95% CI is [0.188, 0.638]

```
# b)
sig<-fit.weib$scale
mu<-fit.weib$coef
phi <- log(sig)
var1 <- matrix(c(-exp(-phi), -(log(6000)-mu)*-exp(-phi)))
var_loglog <- t(var1) %*% fit.weib$var %*% var1
loglog_st <- (log(6000)-mu)/sig
CIlogL <- loglog_st - 1.96*sqrt(var_loglog)
CIlogU <- loglog_st + 1.96*sqrt(var_loglog)
CIU <- exp(-exp(CIlogL))
CIL <- exp(-exp(CIlogU))
CI <- c(CIL, CIU)
CI
```

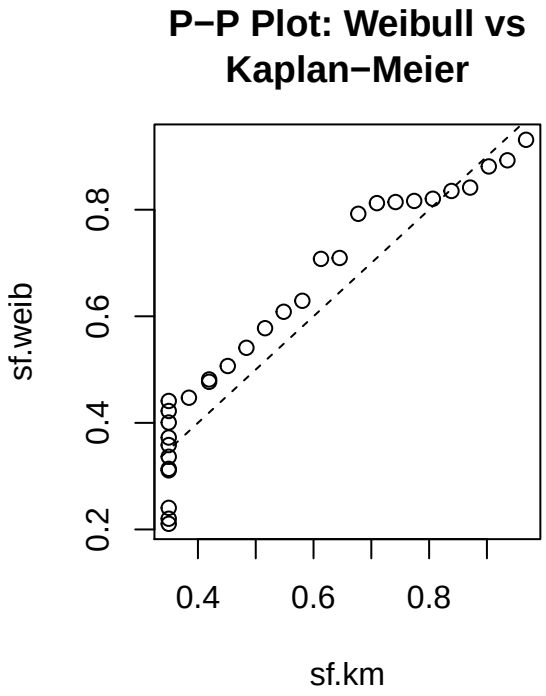
```
## [1] 0.1880582 0.6381961
```

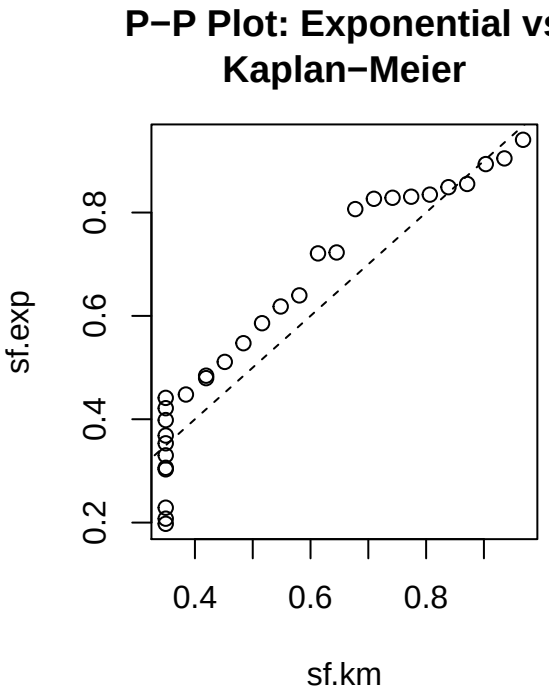
Question 2

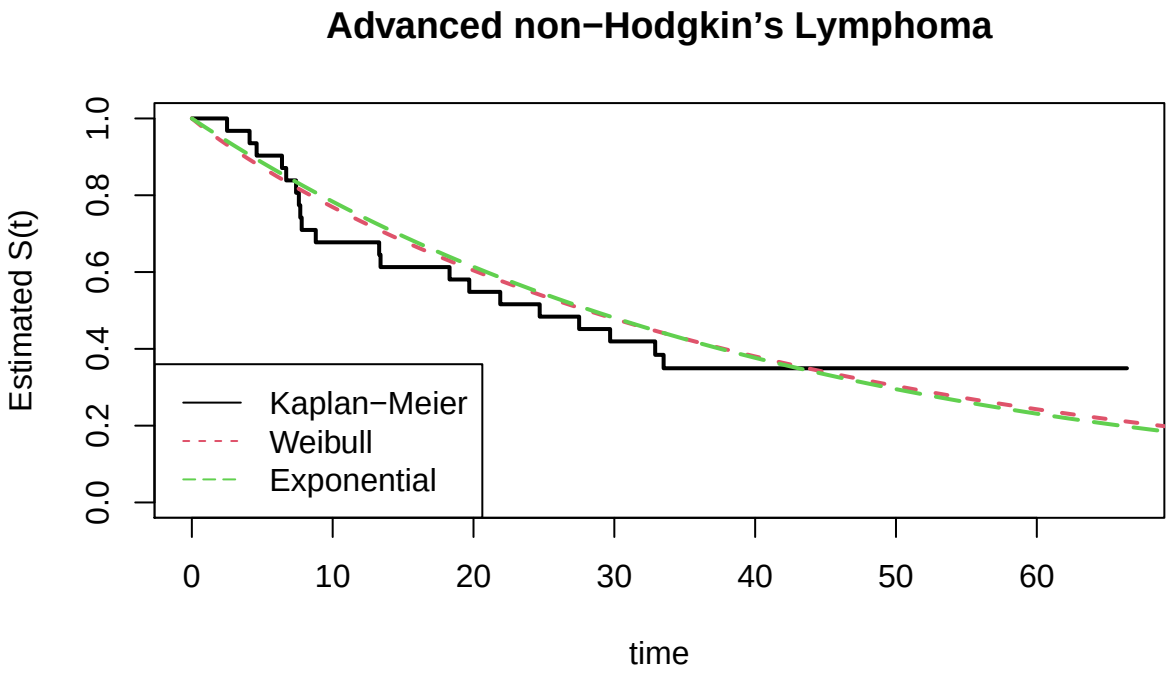
a) If weibull distribution fits well, then the Q-Q plot for $log(-log(Skm(t)))vslog(t)$ should be show an approximatly linear relationship.However, form the plot below, we can see the Q-Q plot is not very linear, this means weibull distribution may not fit the data very well.



b) From the P-P plot, the difference between weibull and exponential estimation are very small and all patterns are similar. From the third plot, we can also see the survival curve for weibull and exponential distribution are very close to each other. However, we can see none of the P-P plot shows a linear pattern and we can also see neither weibull nor exponential distribution can give very accurate estimation between t=5 to t=40. Therefore, exponential distribution is not good enough for the data.

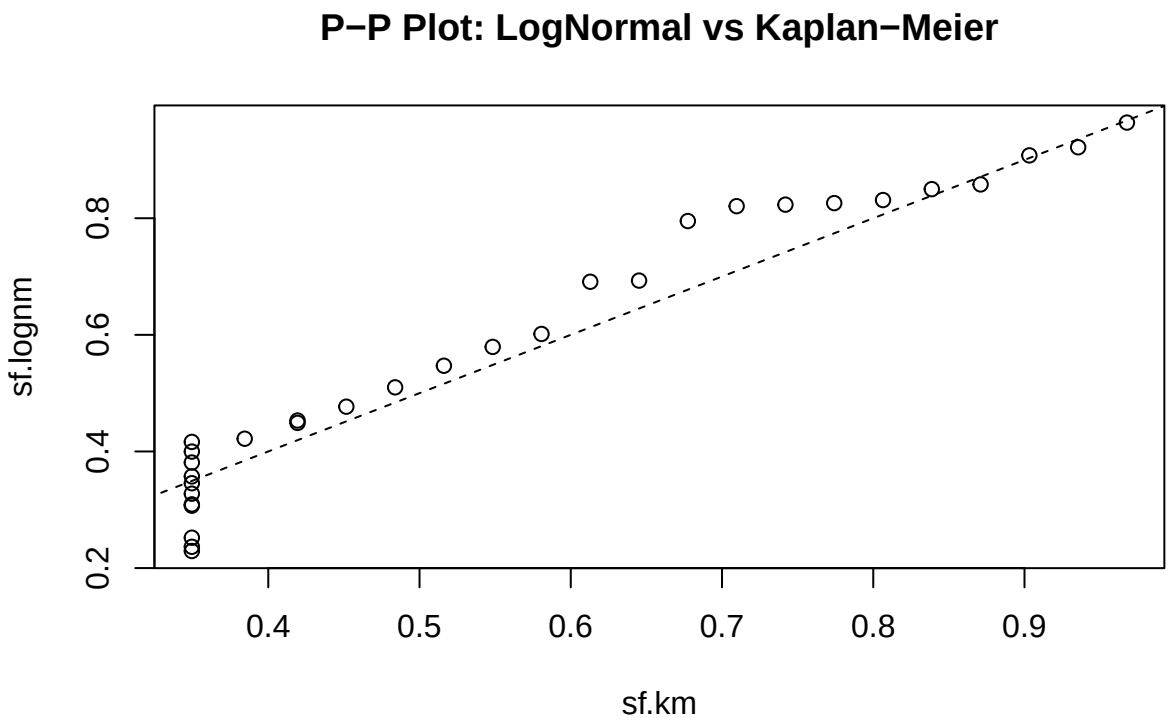


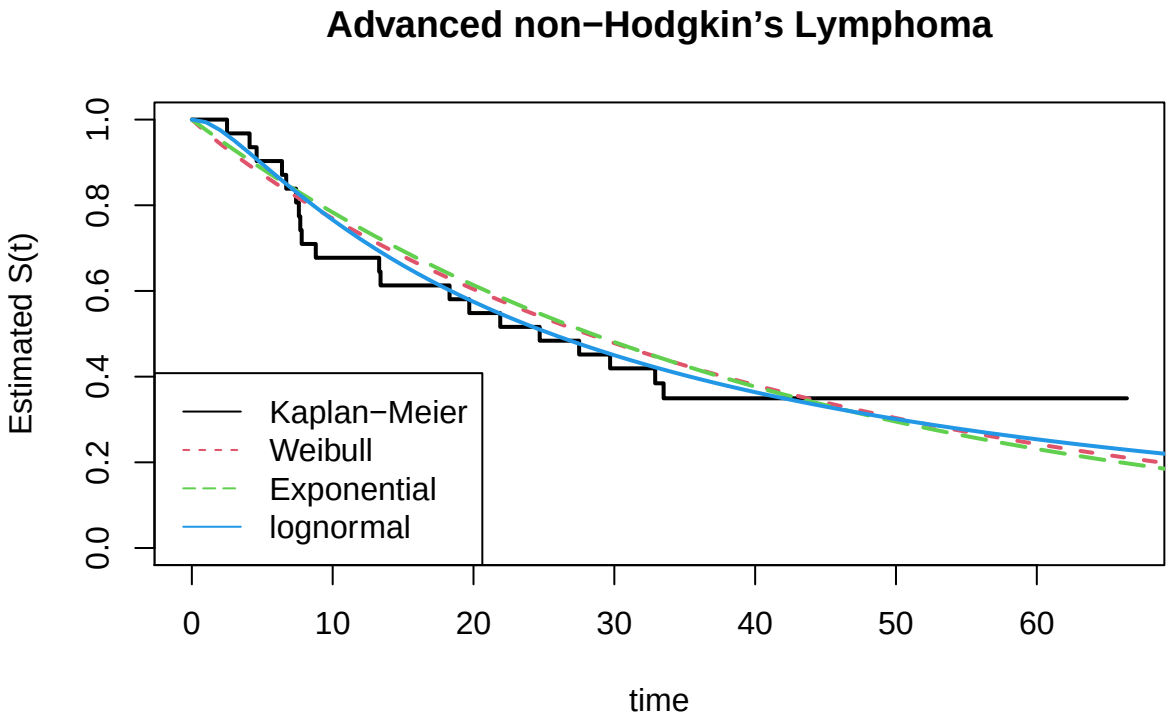




c) From the P-P plot, we can see the points are more closer to the line. therefore, even though the data itself is not perfect and the P-P plot does not show a clear linear pattern, the log normal distribution fit is still better than weibull and exponential distribution. From the second plot, we can clear see the log normal distribution survival curve is more closer to the KM curve compare to exponential and weibull distribution. To conclude, Weibull and exponential model fits the data reasonably well for small and

large survival times, but they do not fit so well for intermediate survival times. log normal model fits the data better in the entire range.





Question 3

a) To test equality of failure time distributions at stress levels 850 and 950 N/mm^2, we need to test $S_1(t) = S_2(t)$. The first step is to test $H_0 : \sigma_1 = \sigma_2$, becasue R code will only provide ϕ, so wwe need to check if $H_0 : \phi_1 = \phi_2$. To test this we need to fit a all model and two separate model for two levels. then we can get the loglike(model) value, add this value together for 859 and 950 levels and minus the all model, we get 0.34 with a p-value 0.31. this means we can not reject the H_0, that is $\sigma_1 = \sigma_2$. The next step is to check $H_0 : \mu_1 = \mu_2$. becasue we already have the model, so we can just use $loglike(model) - loglike(intercept)$ to get teh observed value which equals to 33.68. the p-value is 6.5e-09, this can be obtained from the all model. because 6.5e-09 < 0.05, so we can reject the null hypothesis. this means $\mu_1 ! = \mu_2$. The test suggest that for describing the survival times of the 2 groups, we should consider a Weibull model with $\sigma_1 = \sigma_2$, but $\mu_1 ! = \mu_2$.

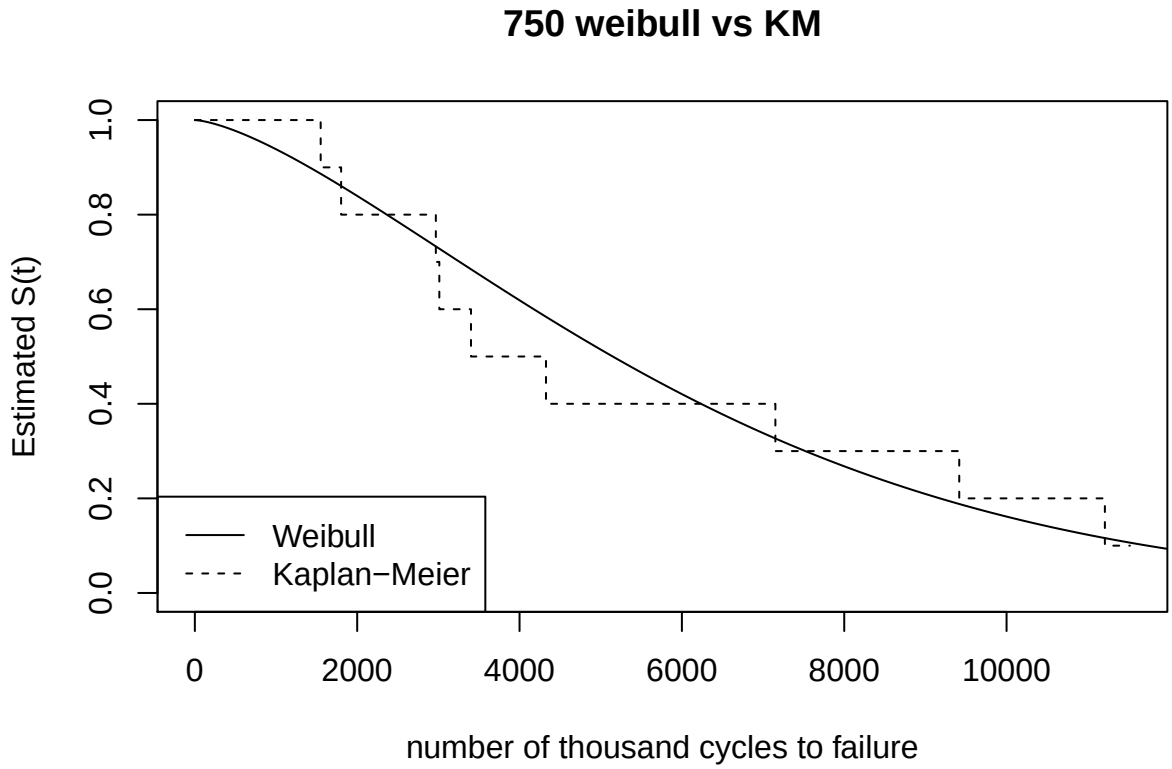
```
##
## Call:
## survreg(formula = Surv(time, status) ~ as.factor(stress), data = data1)
##                          Value Std. Error     z      p
## (Intercept)           5.9111     0.0492 120.3 <2e-16
## as.factor(stress)950 -0.7033     0.0683 -10.3 <2e-16
## Log(scale)           -1.8812     0.1746 -10.8 <2e-16
##
## Scale= 0.152
##
## Weibull distribution
## Loglik(model)= -103.1   Loglik(intercept only)= -120
##  Chisq= 33.68 on 1 degrees of freedom, p= 6.5e-09
## Number of Newton-Raphson Iterations: 6
## n= 20
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = data1[data1$stress ==
##     850, ])
##               Value Std. Error      z       p
## (Intercept)  5.9176     0.0454 130.35 < 2e-16
## Log(scale)  -1.9951     0.2510  -7.95 1.9e-15
##
## Scale= 0.136
##
## Weibull distribution
## Loglik(model)= -54.2   Loglik(intercept only)= -54.2
## Number of Newton-Raphson Iterations: 6
## n= 10
```

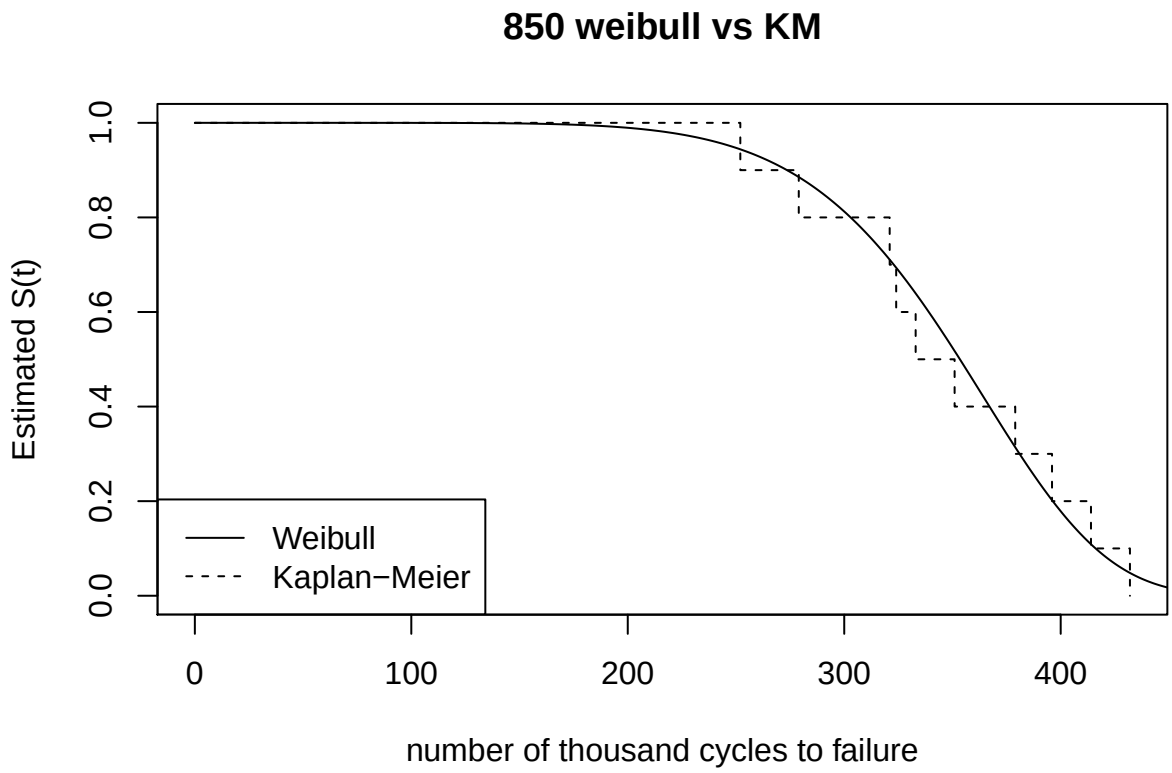
```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = data1[data1$stress ==
##     950, ])
##               Value Std. Error     z       p
## (Intercept)  5.2007     0.0559 92.98 < 2e-16
## Log(scale)  -1.7879     0.2442 -7.32 2.5e-13
##
## Scale= 0.167
```

```
##
## Weibull distribution
## Loglik(model)= -48.8   Loglik(intercept only)= -48.8
## Number of Newton-Raphson Iterations: 6
## n= 10

## Observed lambda value is 0.3486495

## p-value =  0.3173105

## Observed mu value is 33.67751

## p-value = 6.5e-09
```

b) The first thing we need to do is fit Weibull and KM model into the three stress level. Then from the plots below, we can see Weibull distribution fits the data pretty good for all three stress level, the overall pattern was similar to the KM curve. We can also see level 850 and 950 has more similar pattern. In level 750, the Weibull fit was slightly overestimated between t=3000 to 6000. We can also see the time fpr level 750 is much larger than 850 and 950 group. Therefore, as the stress level increases, the failure time would be much shorter. To get the location and scale value, we can just use fit$coef and fit$scale functions for level 750 and 850, for level 950, we need to use the summary from fitall model, the location value is fitall$coeff[2]+mu1, and the scale value is the same as level 850. Therefoe, we have the value for level 750 is location = 8.798, scale = 0.686. for level 850 the location = 5.911, scale = 0.152, for level 950, the location = 5.911106-0.7032717 which is 5.208, scale = 0.152

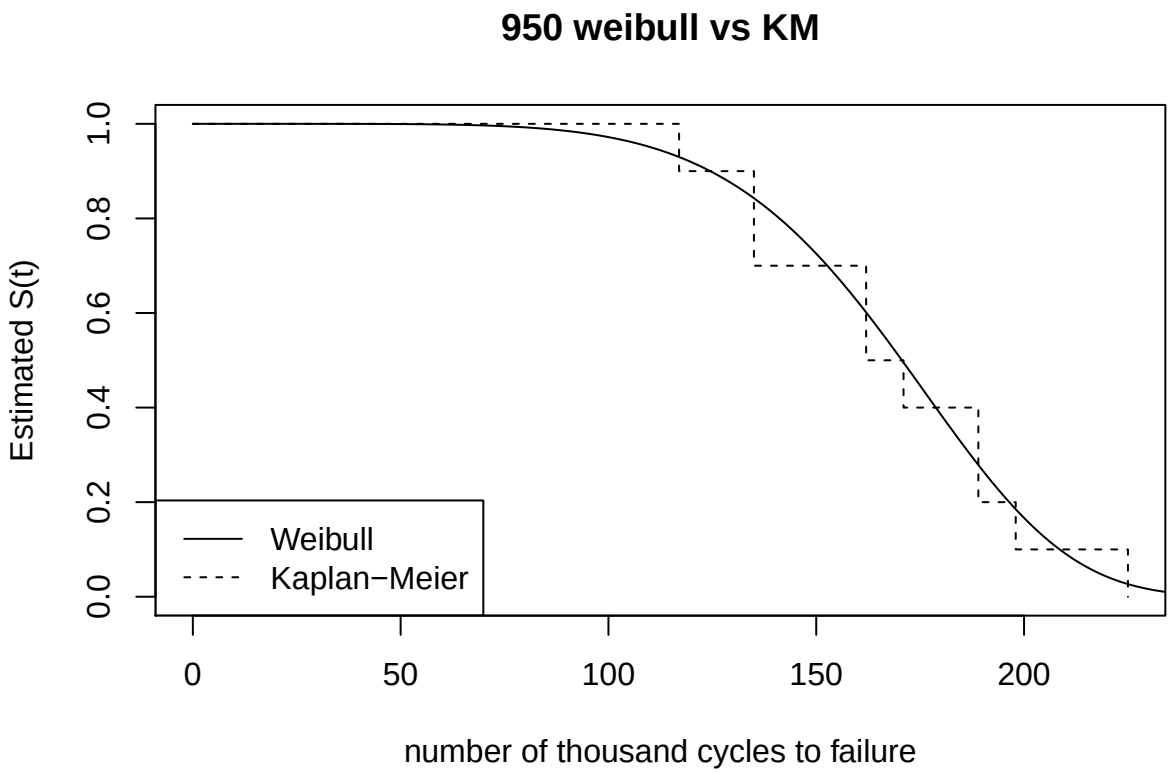


```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = data[data$stress ==
##     750, ])
##              Value Std. Error     z      p
## (Intercept)  8.798      0.234 37.55 <2e-16
## Log(scale)  -0.376      0.267 -1.41   0.16
##
## Scale= 0.686
##
## Weibull distribution
## Loglik(model)= -86.8   Loglik(intercept only)= -86.8
## Number of Newton-Raphson Iterations: 6
## n= 10
```




```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = data1[data1$stress ==
##     850, ])
##               Value Std. Error      z       p
## (Intercept)  5.9176     0.0454 130.35 < 2e-16
## Log(scale)  -1.9951     0.2510  -7.95 1.9e-15
##
## Scale= 0.136
##
## Weibull distribution
```

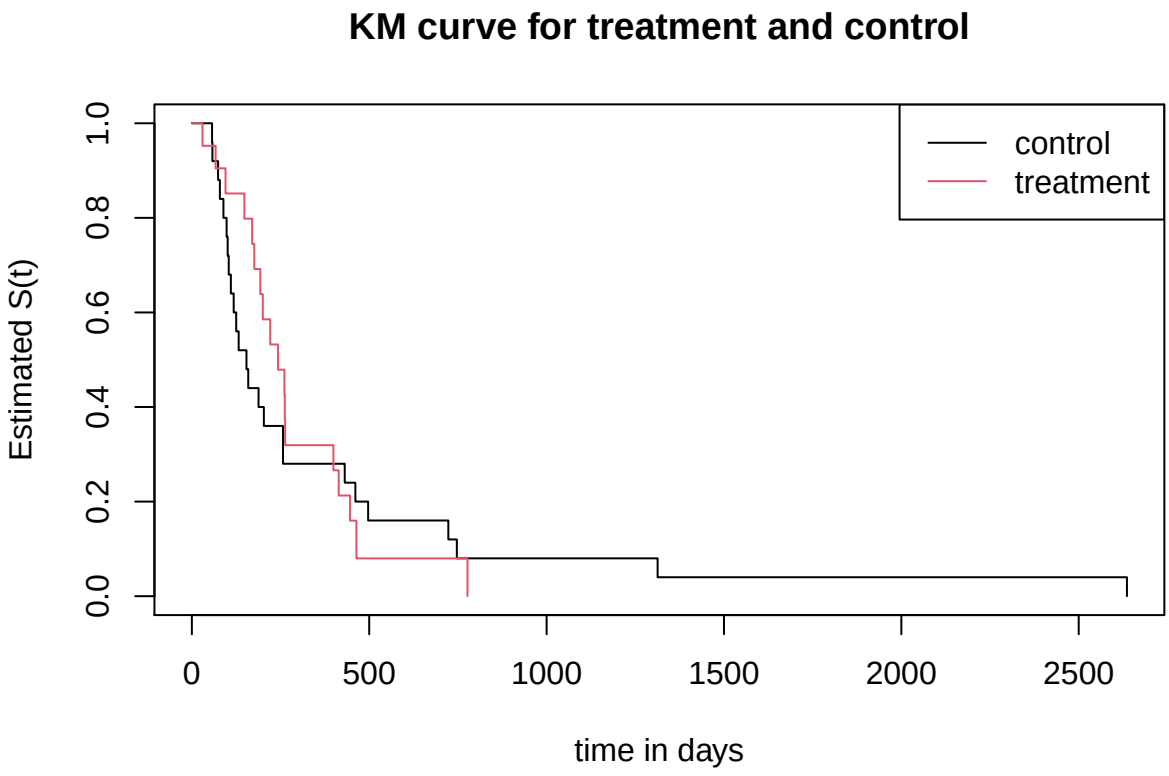
```
## Loglik(model)= -54.2   Loglik(intercept only)= -54.2
## Number of Newton-Raphson Iterations: 6
## n= 10
```




```
##
## Call:
## survreg(formula = Surv(time, status) ~ 1, data = data1[data1$stress ==
##     950, ])
##               Value Std. Error     z       p
## (Intercept)  5.2007     0.0559 92.98 < 2e-16
## Log(scale)  -1.7879     0.2442 -7.32 2.5e-13
##
## Scale= 0.167
##
## Weibull distribution
## Loglik(model)= -48.8   Loglik(intercept only)= -48.8
## Number of Newton-Raphson Iterations: 6
## n= 10

## The location parameter for level 950 is  5.911106 -0.7032717
```

Question 4

a) From the plot below, it seems like the control group has longer survival time, this means the treatment will acctually decrease the survival time in long term. We can also see when the survival time is between t=100 and t=400, the treatment group has hiogher survial rate, control group has higher survival rate after t=400 and before t= 100. This suggest that this radiation drug theraoy will provide survival benefit in short term but not longterm.

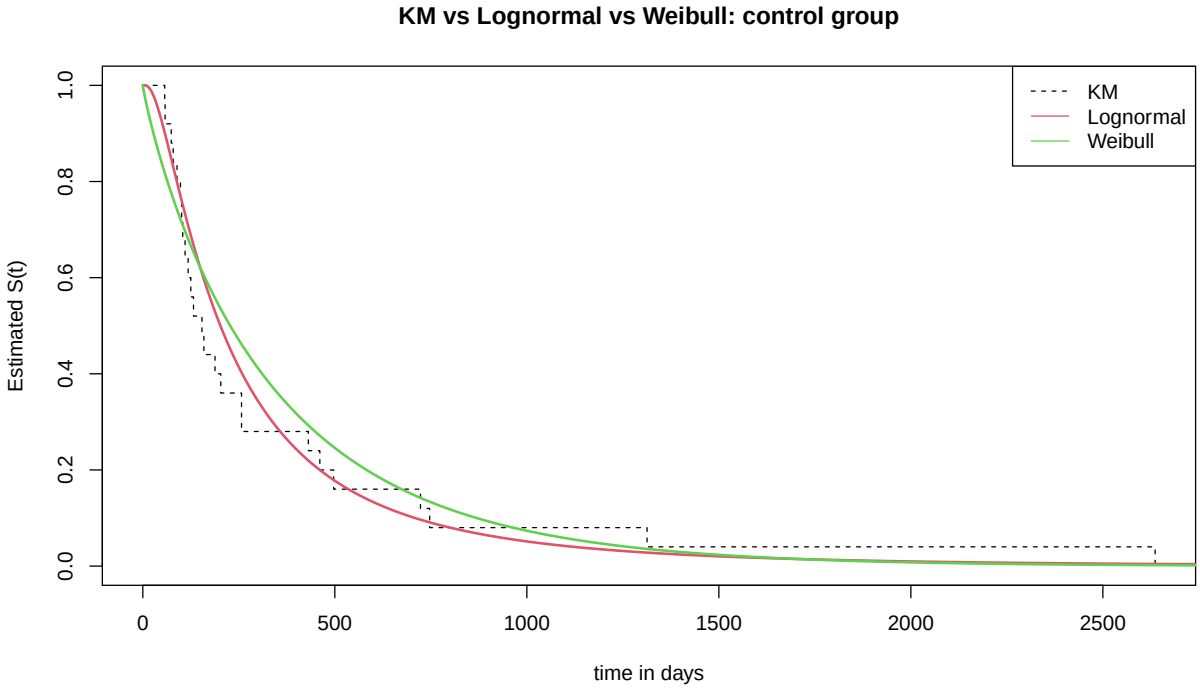


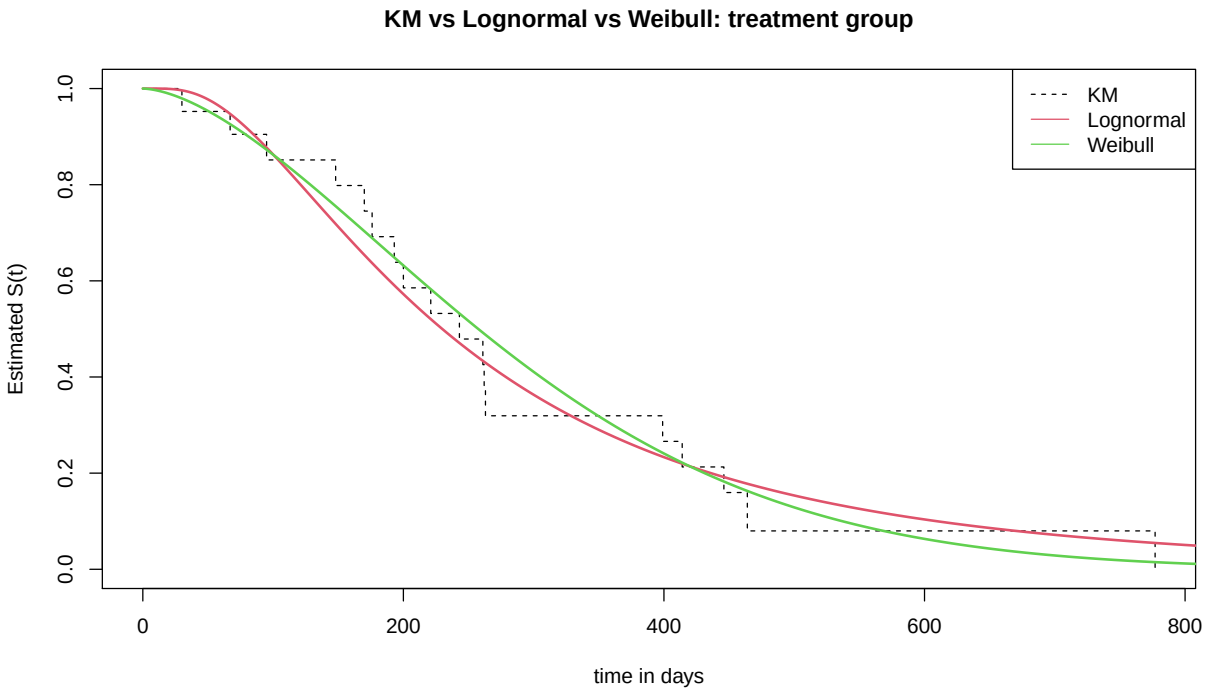
b) From the Logrank test, we can see the p-value is 0.8 which is greater than 0.05, this means we can not reject eh null hypothesis, and treatment S(t) = control S(t). From the R code result, we know the two group have sam efailure time distribution. From the willcoxon test, we can see the p-value is 0.3 which is also greater than 0.05. this means we can not reject the null hypothesis, and treatment S(t) = control S(t).From the R code result, we know the two group have sam efailure time distribution. In the above plot, we can see there is cross over occur, this means none of the logrank nor wilcoxon tests are suitable. Weight log rank test are meant to detect difference between s.f.s when the 2 groups do not cross over in survival function. because the difference cancels out.

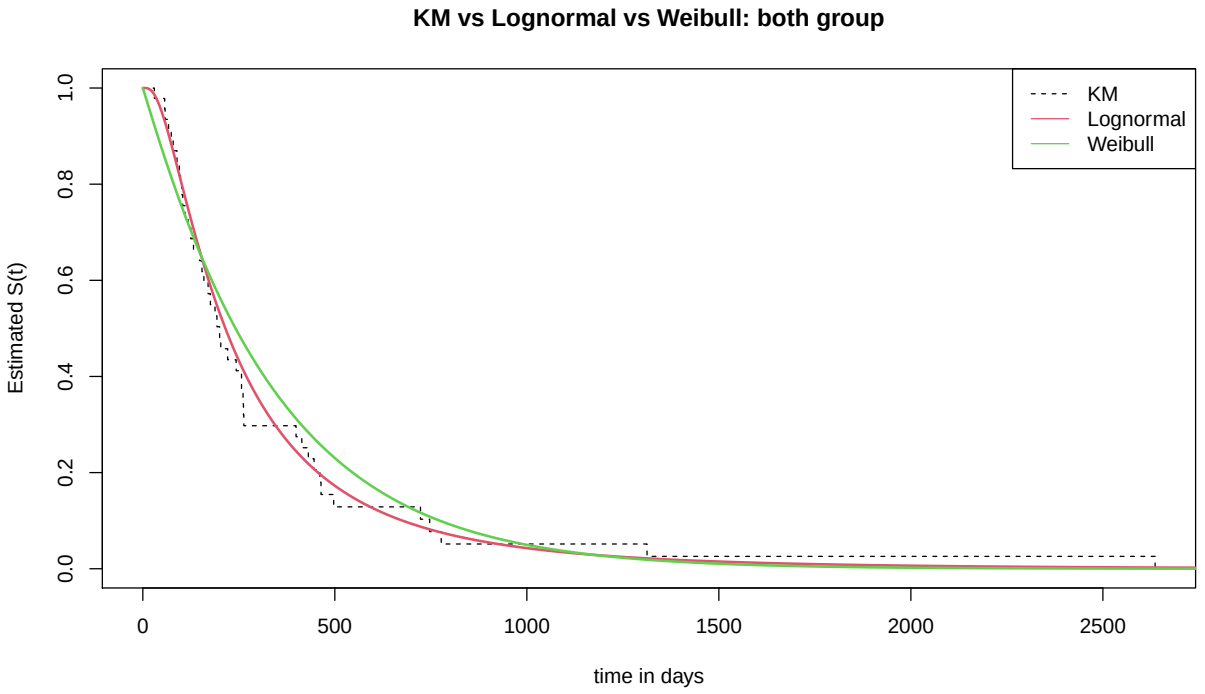
```
## Call:
## survdiff(formula = Surv(time, status) ~ as.factor(trt), data = bdc)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## as.factor(trt)=0 25       25     24.2    0.0268    0.0667
## as.factor(trt)=1 21       18     18.8    0.0345    0.0667
##
##  Chisq= 0.1  on 1 degrees of freedom, p= 0.8

## Call:
```

```
## survdiff(formula = Surv(time, status) ~ as.factor(trt), data = bdc,
##     rho = 1)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## as.factor(trt)=0 25    13.93     11.7     0.423      1.31
## as.factor(trt)=1 21     8.62     10.8     0.457      1.31
##
##  Chisq= 1.3  on 1 degrees of freedom, p= 0.3
```

c) The most commen parametric methods are log-normal and weibull, so we will test for both method in here. To see the fit for each method, we need to plot KM and the fited curve together to see the difference. From the control group plot, we can see lognormal distribution fits better becasue the Weibull model overestimate the survival rate in the most of the time point, lognormal model has more closer value to the km curve. In the treatment group, we can see Weibull model fits better in the early and late time points since the lognormal model will underestimate the suirival rate early and overestimate at the end. The lognrmal model fits better in the middle. To see which one is better, we still need a plot to check the overall fit. From the third plot, we can see lognormal distribution fits better, so we can conclude that the lognormal model performs better overall. Thus, the lognormal model is appropriate for the data. Further more, becasue the sample size is quite small, the model fit maybe different as the sample size increases.




KM vs Lognormal vs Weibull: treatment group
KM
Lognormal
Weibull
Estimated S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
time in days


KM vs Lognormal vs Weibull: both group
KM
Lognormal
Weibull
Estimated S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
time in days

Question 5

Appendix

Question 1

```
library(survival)
data<-read.table("p31.txt", header=T)
data <- data[data$stress == 750,]
x<-data$time
delta<-data$status
fit.weib<- survreg(Surv(x, delta)~1)
summary(fit.weib)

y <- log(-log(0.5))
y_mid <- fit.weib$coef + y*fit.weib$scale
vary_mid <- as.numeric(t(matrix(c(1, y*fit.weib$scale))) %*%
                         as.matrix(fit.weib$var) %*%
                         matrix(c(1, y*fit.weib$scale)))
CIL <- exp(y_mid - 1.96*sqrt(vary_mid))
CIU <- exp(y_mid + 1.96*sqrt(vary_mid))
CI <- c(CIL, CIU)
CI
```

```
# b)
sig<-fit.weib$scale
mu<-fit.weib$coef
phi <- log(sig)
var1 <- matrix(c(-exp(-phi), -(log(6000)-mu)*-exp(-phi)))
var_loglog <- t(var1) %*% fit.weib$var %*% var1
loglog_st <- (log(6000)-mu)/sig
CIlogL <- loglog_st - 1.96*sqrt(var_loglog)
CIlogU <- loglog_st + 1.96*sqrt(var_loglog)
CIU <- exp(-exp(CIlogL))
CIL <- exp(-exp(CIlogU))
CI <- c(CIL, CIU)
CI
```

Question 2

```
time <- c(2.5, 4.1, 4.6, 6.4, 6.7, 7.4, 7.6, 7.7, 7.8, 8.8, 13.3, 13.4, 18.3,
          19.7, 21.9, 24.7, 27.5, 29.7, 30.1, 32.9, 33.5, 35.4, 37.7, 40.9,
          42.6, 45.4, 48.5, 48.9, 60.4, 64.4, 66.4)
status <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 0)
data <- as.data.frame(cbind(time,status))
fit <- survfit(Surv(time, status)~1)
llt <- log(-log(fit$surv))
lt <- log(time)
plot(lt,llt,main="Q-Q Plot: log(-log(Skm(t))) vs
log(t)")
lines(seq(0,7,0.1),seq(-3.5,3.5,0.1), lty=2)
```

```
fit.weib <- survreg(Surv(time, status)~1, data = data)
sig1<-fit.weib$scale
mu1<-fit.weib$coef
time<-fit$time
sf.weib<-exp(-exp((log(time)-mu1)/sig1))

fit.exp <- survreg(Surv(time, status)~1, dist = "exp", data = data)
mu2 <- fit.exp$coef
sf.exp <- exp(-exp((log(time)-mu2)))

sf.km<-fit$surv

par(mfrow=c(1,2))
plot(sf.km,sf.weib,main="P-P Plot: Weibull vs
Kaplan-Meier")
lines(seq(0,1,0.1),seq(0,1,0.1), lty=2)
plot(sf.km, sf.exp, main="P-P Plot: Exponential vs
Kaplan-Meier")
lines(seq(0,1,0.1),seq(0,1,0.1), lty=2)

par(mfrow=c(1,1))
t<-0:175
st.weib<-exp(-exp((log(t)-mu1)/sig1))
st.exp<-exp(-exp(log(t)-mu2))
plot(fit,xlab="time",
     ylab="Estimated S(t)",main="Advanced non-Hodgkin's Lymphoma",
     conf.int=F,lty=1,lwd=2)
lines(t,st.weib,lty=2,lwd=2, col=2)
lines(t,st.exp,lty=5,lwd=2, col=3)
legend("bottomleft", c("Kaplan-Meier", "Weibull", "Exponential"),
       col=c(1,2,3), lty=c(1,2,5))
```

```
par(mfrow=c(2,1))

fit.lognm<-survreg(Surv(time,status)~1,data=data, dist="lognormal")
mu3<-fit.lognm$coeff
sig3<-fit.lognm$scale
sf.lognm<-1-pnorm((log(time)-mu3)/sig3)
plot(sf.km, sf.lognm, main="P-P Plot: LogNormal vs Kaplan-Meier")
lines(seq(0,1,0.1),seq(0,1,0.1), lty=2)

t<-0:175
st.weib<-exp(-exp((log(t)-mu1)/sig1))
st.exp<-exp(-exp(log(t)-mu2))
st.lognm<-1-pnorm((log(t)-mu3)/sig3)
plot(fit,xlab="time",
     ylab="Estimated S(t)",main="Advanced non-Hodgkin's Lymphoma",
     conf.int=F,lty=1,lwd=2)
lines(t,st.weib,lty=2,lwd=2, col=2)
lines(t,st.exp,lty=5,lwd=2, col=3)
lines(t,st.lognm,lty=7,lwd=2, col=4)
legend("bottomleft", c("Kaplan-Meier", "Weibull", "Exponential", "lognormal"),
       col=c(1,2,3,4), lty=c(1,2,5,7))
```

Question 3

```
data<-read.table("p31.txt", header=T)
data1 <- data[data$stress != 750,]
fitall <- survreg(Surv(time, status)~as.factor(stress), data = data1)
fit0 <- survreg(Surv(time, status)~1, data = data[data$stress == 750,])
fit1 <- survreg(Surv(time, status)~1, data = data1[data1$stress == 850,])
fit2 <- survreg(Surv(time, status)~1, data = data1[data1$stress == 950,])
print(summary(fitall))
print(summary(fit1))
print(summary(fit2))
# Test on phi
lambda_obs <- 2*(fit1$loglik[2] + fit2$loglik[2] - fitall$loglik[2])
cat("Observed lambda value is", lambda_obs, "\n")
sigmap <- pchisq(1,df=1, lower.tail = F)
cat("p-value = ", sigmap, "\n")
# Test on mu
mu_obs <- 2*(fitall$loglik[2] - fitall$loglik[1])
cat("Observed mu value is", mu_obs, "\n")
cat("p-value = 6.5e-09 ")
```

```
# for 750 level
sig <- fit0$scale
mu <- fit0$coef
fit_km <- survfit(Surv(time, status)~1, data = data[data$stress == 750,])
t <- 0:12000
sf <- exp(-exp((log(t)-mu)/sig))
plot(fit_km, xlab="number of thousand cycles to failure",
             ylab = "Estimated S(t)", main = "750 weibull vs KM",
     conf.int = F, lty = 2)
lines(t, sf, lty = 1)
legend("bottomleft", c("Weibull", "Kaplan-Meier"), lty = c(1,2))
summary(fit0)

# for 850 level
sig <- fit1$scale
mu <- fit1$coef
fit_km <- survfit(Surv(time, status)~1, data = data1[data1$stress == 850,])
t <- 0:12000
sf <- exp(-exp((log(t)-mu)/sig))
plot(fit_km, xlab="number of thousand cycles to failure",
             ylab = "Estimated S(t)", main = "850 weibull vs KM",
     conf.int = F, lty = 2)
lines(t, sf, lty = 1)
legend("bottomleft", c("Weibull", "Kaplan-Meier"), lty = c(1,2))
summary(fit1)

# for 950 level
sig <- fit2$scale
mu <- fit2$coef
fit_km <- survfit(Surv(time, status)~1, data = data1[data1$stress == 950,])
t <- 0:12000
sf <- exp(-exp((log(t)-mu)/sig))
```

```
plot(fit_km, xlab="number of thousand cycles to failure",
             ylab = "Estimated S(t)", main = "950 weibull vs KM",
     conf.int = F, lty = 2)
lines(t, sf, lty = 1)
legend("bottomleft", c("Weibull", "Kaplan-Meier"), lty = c(1,2))
summary(fit2)

# location for level 950
x <- fitall$coefficients
cat("The location parameter for level 950 is ", x)
```

Question 4

```
bdc <- read.table("p34.txt", header=T)
fit.km <- survfit(Surv(time, status)~as.factor(trt), data = bdc, conf.int=F)
plot(fit.km, xlab = "time in days", ylab = "Estimated S(t)",
     main= "KM curve for treatment and control", col = 1:2)
legend("topright", c("control", "treatment"),col = 1:2, lty = 1:1)
```

```
logrank <- survdiff(Surv(time, status)~as.factor(trt), data = bdc)
logrank
wilcox <- survdiff(Surv(time, status)~as.factor(trt), data = bdc, rho = 1)
wilcox
```

```
# for control and treatment group separately
ctrl_km <- survfit(Surv(time, status)~as.factor(trt),
                   data = bdc[bdc$trt == 0,], conf.int=F)
trt_km <- survfit(Surv(time, status)~as.factor(trt),
                  data = bdc[bdc$trt == 1,],  conf.int=F)
t <- 0:3000
# lognormal control
lctrl <- survreg(Surv(time, status)~1, data = bdc[bdc$trt == 0,],
                 dist = "lognormal")
mulc <- lctrl$coefficients
siglc <- lctrl$scale
stlc <- 1 - pnorm((log(t) - mulc)/siglc)
# lognormal treatment
ltrt <- survreg(Surv(time, status)~1, data = bdc[bdc$trt == 1,],
                dist = "lognormal")
mult <- ltrt$coefficients
siglt <- ltrt$scale
stlt <- 1 - pnorm((log(t) - mult)/siglt)
# Weibull control
wctrl <- survreg(Surv(time, status)~1, data = bdc[bdc$trt == 0,])
muwc <- wctrl$coefficients
sigwc <- wctrl$scale
stwc <- exp(-exp((log(t)-muwc)/sigwc))
# Weibull treatment
wtrt <- survreg(Surv(time, status)~1, data = bdc[bdc$trt == 1,])
muwt <- wtrt$coefficients
sigwt <- wtrt$scale
```

```
stwt <- exp(-exp((log(t)-muwt)/sigwt))

plot(ctrl_km, xlab = "time in days", ylab = "Estimated S(t)",
     main= "KM vs Lognormal vs Weibull: control group", conf.int = F, lty = 2)
lines(t, stlc, lty = 1, lwd = 2, col = 2)
lines(t, stwc, lty = 1, lwd = 2, col = 3)
legend("topright", c("KM", "Lognormal", "Weibull"), col = c(1,2,3), lty = c(2,1,1))

plot(trt_km, xlab = "time in days", ylab = "Estimated S(t)",
     main= "KM vs Lognormal vs Weibull: treatment group", conf.int = F, lty = 2)
lines(t, stlt, lty = 1, lwd = 2, col = 2)
lines(t, stwt, lty = 1, lwd = 2, col = 3)
legend("topright", c("KM", "Lognormal", "Weibull"), col = c(1,2,3), lty = c(2,1,1))

#overall

all_km <- survfit(Surv(time, status)~1,
                  data = bdc, conf.int=F)
wall <- survreg(Surv(time, status)~1, data = bdc)
muw <- wall$coefficients
sigw <- wall$scale
stw <- exp(-exp((log(t)-muw)/sigw))
lall <- survreg(Surv(time, status)~1, data = bdc, dist = "lognormal")
mul <- lall$coefficients
sigl <- lall$scale
stl <- 1 - pnorm((log(t) - mul)/sigl)

plot(all_km, xlab = "time in days", ylab = "Estimated S(t)",
     main= "KM vs Lognormal vs Weibull: both group", conf.int = F, lty = 2)
lines(t, stl, lty = 1, lwd = 2, col = 2)
lines(t, stw, lty = 1, lwd = 2, col = 3)
legend("topright", c("KM", "Lognormal", "Weibull"), col = c(1,2,3), lty = c(2,1,1))
```