STAT5703 HW2 Ex1

Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

##Exercise 1.

```r
library(readr)
cancer <- read_table2("cancer.txt")[,2:7]
```

```
## Parsed with column specification:
## cols(
##   index = col_double(),
##   day = col_double(),
##   month = col_double(),
##   year = col_double(),
##   t = col_double(),
##   d = col_double(),
##   arm = col_character()
## )
```

####1. *Answer* : It is reasonable to say the right censoring in this data set is random because the patients who are right censored(with d=0) have various censoring time. So censoring time is not fixed.

####2.

```r
library(ggfortify)
```

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

```r
library(survival)
```

```
##
## Attaching package: 'survival'

## The following object is masked _by_ '.GlobalEnv':
##
##     cancer
```

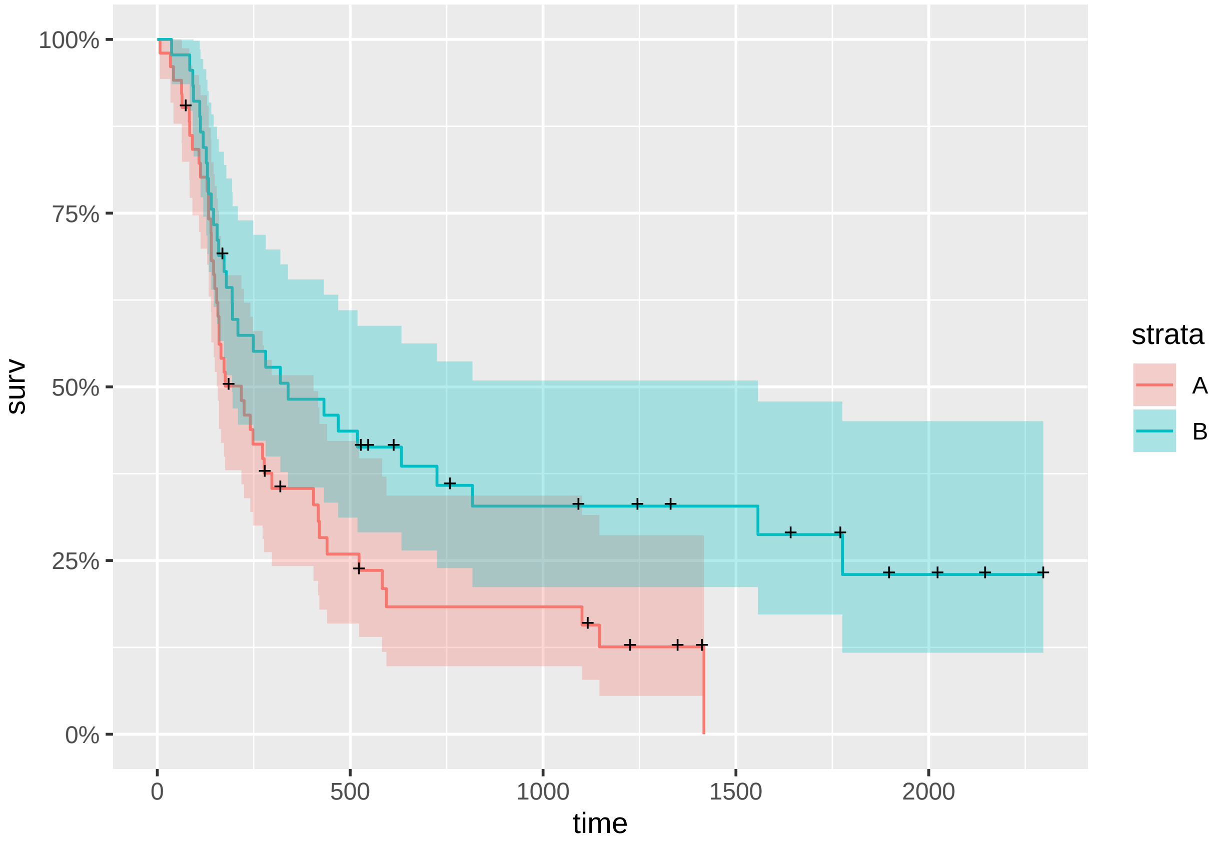
```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
autoplot(survfit(Surv(t,d)~arm,data=cancer),conf.int=TRUE,col=c(2,3),title="Kaplan-Meier estimates for 
```



* The plot seems to indicate a difference between two groups. Treatment B seems to be more efficient.

####3.

```
cancer$arm<-factor(cancer$arm)
cancer0<-cancer%>%filter(cancer$arm=="A")
cancer1<-cancer%>%filter(cancer$arm=="B")
fit_exp<-survreg(Surv(t,d)~arm,data = cancer,dist = "exponential")
summary(fit_exp)
```

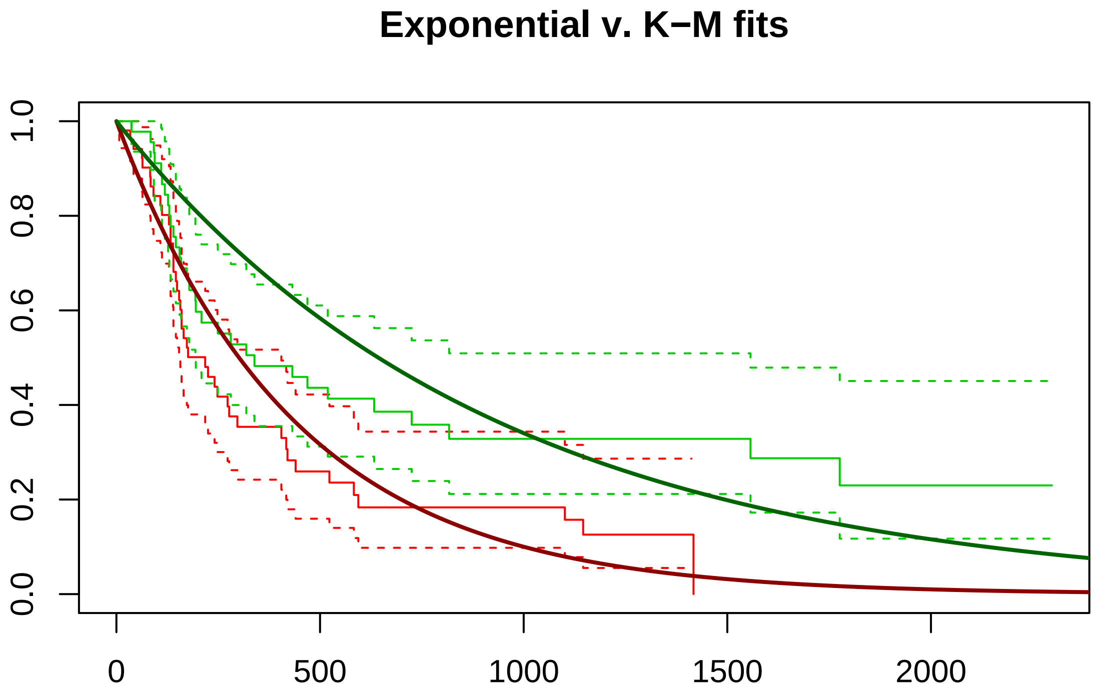
```
##
## Call:
## survreg(formula = Surv(t, d) ~ arm, data = cancer, dist = "exponential")
##             Value Std. Error    z      p
## (Intercept) 6.074      0.154 39.4 <2e-16
## armB        0.759      0.237  3.2 0.0014
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -539.9   Loglik(intercept only)= -545.1
##  Chisq= 10.41 on 1 degrees of freedom, p= 0.0013
## Number of Newton-Raphson Iterations: 4
## n= 96
```

- From the summary table, we can see that the estimated value of armB is 0.759>0. So we can know that the treatment B is efficient and can increase the survival time. It agrees with my intuition from last point.

####4. * From the summary table in point 3, we get Chisq=10.41 which is the likelihood ratio test between two groups. Its p value is 0.0013 so there is strong evidence that the model which includes both armA and armB is better than intercept only and thus the effect of treatment B is significant. This conclusion does not depend on parametric model assumptions

####5.

```
plot(survfit(Surv(t,d)~arm,data=cancer),conf.int=TRUE,col=c(2,3),main="Exponential v. K-M fits")
x <- seq(from=0,to=2500,by=1)
lines(x,1-pexp(x,exp(-coef(fit_exp)[1])),col="darkred",lwd=2)
lines(x,1-pexp(x,exp(-sum(coef(fit_exp)))),col="darkgreen",lwd=2)
```

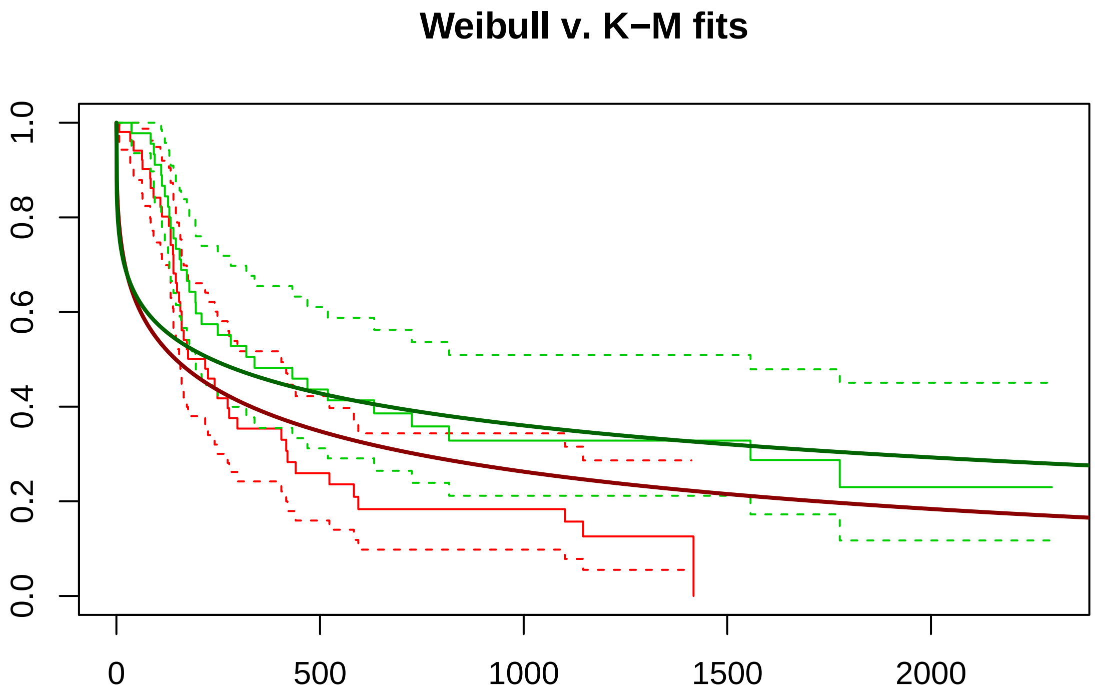


* From the graph, we can see that exponential model does great to fit. The lines lie inside the corresponding group's confidence interval. It does extremely good for fitting group A and it correctly goes to zero after group A is not cencored. But it is not really good when fitting group B as it the fitted line goes outside of confidence interval after around time=1500 and it also goes to zero even though group B is still right censored.

####6.

```
fit_wei12<-survreg(Surv(t,d)~arm,data = cancer0,dist = "weibull")
fit_wei22<-survreg(Surv(t,d)~arm,data = cancer1,dist = "weibull")
gamma12=1/exp(fit_wei12$scale)
```

```
gamma22=1/exp(fit_wei22$scale)
plot(survfit(Surv(t,d)~arm,data=cancer),conf.int=TRUE,col=c(2,3),main="Weibull v. K-M fits")
x <- seq(from=0,to=2500,by=1)
lines(x,1-pweibull(x,gamma12,exp(coef(fit_wei12)[1])),col="darkred",lwd=2)
lines(x,1-pweibull(x,gamma22,exp(coef(fit_wei22)[1])),col="darkgreen",lwd=2)
```



```
1-pchisq(2*(fit_wei12$loglik[2]+fit_wei22$loglik[2]-fit_exp$loglik[2]),df=1)
```

```
## [1] 0.03242579
```

- This model fits the data better by simply looking at the two fitted lines in each graph. The evidence against the relevence on an exponential model is obtained by calculating likelihood ratio test between Weibull and exponential models. Then we can get a p-value=0.0324<0.05. It means the Weibull model is better than the exponential model.