EAS509 Homework 6 (100 points). Key

Submit your answers as a single pdf attach all R code. Failure to do so will result in grade reduction.

Question 1 (50 points)

For each question, state whether or not the censoring mechanism is independent. Justify your answer with a short statement. (10 points for each)

"Independent censoring essentially means that within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience."

Reference: Survival analysis: A self-learning text" by Kleinbaum and Klein (3rd edition, 2011, Springer)

a) In a study of disease relapse, due to a careless research scientist, all patients whose phone numbers begin with the number "2" are lost to follow up.

Independent: As outcome disease relapse outcome is not positive for all patients and is independent. It does not violate the assumption of independent censoring.

b) In a study of longevity, a formatting error causes all patient ages that exceed 99 years to be lost (i.e. we know that those patients are more than 99 years old, but we do not know their exact ages).

Dependent: Longevity analysis depends on all patient data and hence it violates the assumption of independent censoring.

c) Hospital A conducts a study of longevity. However, very sick patients tend to be transferred to Hospital B, and are lost to follow up.

Dependent: The study of longevity depends on the people who are very sick to bring the curve down for real world estimates. Hence, it violates the assumption of independent censoring.

d) In a study of unemployment duration, the people who find work earlier are less motivated to stay in touch with study investigators, and therefore are more likely to be lost to follow up.

Dependent:People who find work might be unemployed later by choice or by situations and might or might not follow up.Hence it violates the assumption of independent censoring.

e) In a study of pregnancy duration, women who deliver their babies pre-term are more likely to do so away from their usual hospital, and thus are more likely to be censored, relative to women who deliver full-term babies.

Dependent: The duration analysis of pregnancy depends on the women who deliver preterm by a huge extend. Hence, it violates the assumption of independent censoring.

Question 2 (50 points)

A data set from "DATA.csv" represents publication times for 244 clinical trials funded by the National Heart, Lung, and Blood Institute. Using Log-Rank Test in R, estimate if the Kaplan-Meier Survival Curves from two subpopulations stratified by "posres" variable are significantly different.

```
library(data.table)
library(ggplot2)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
df = read.csv("Data.csv")
head(df,5)
```

```
##   posres multi clinend     mech sampsize    budget impact      time status
## 1      0     0       1      R01    39876  8.016941 44.016 11.203285      1
## 2      0     0       1      R01    39876  8.016941 23.494 15.178645      1
## 3      0     0       1      R01     8171  7.612606  8.391 24.410678      1
## 4      0     0       1 Contract    24335 11.771928 15.402  2.595483      1
## 5      0     0       1 Contract    33357 76.517537 16.783  8.607803      1
```

```
str(df$posres)
```

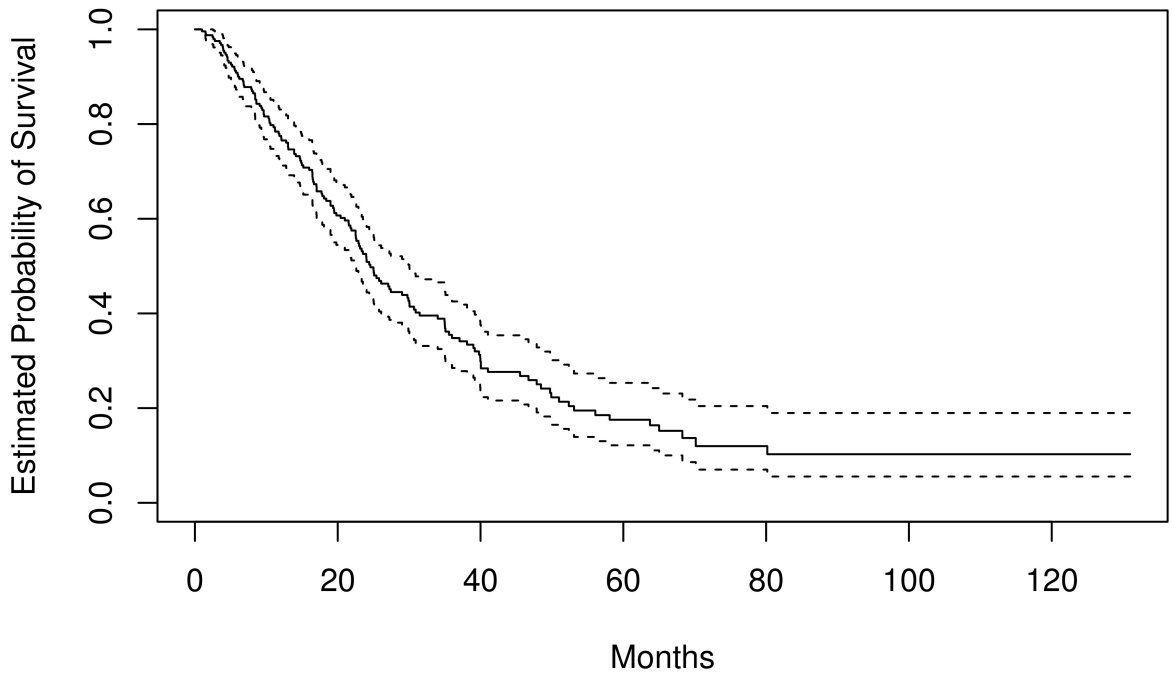
```
##  int [1:244] 0 0 0 0 0 0 0 1 0 0 ...
```

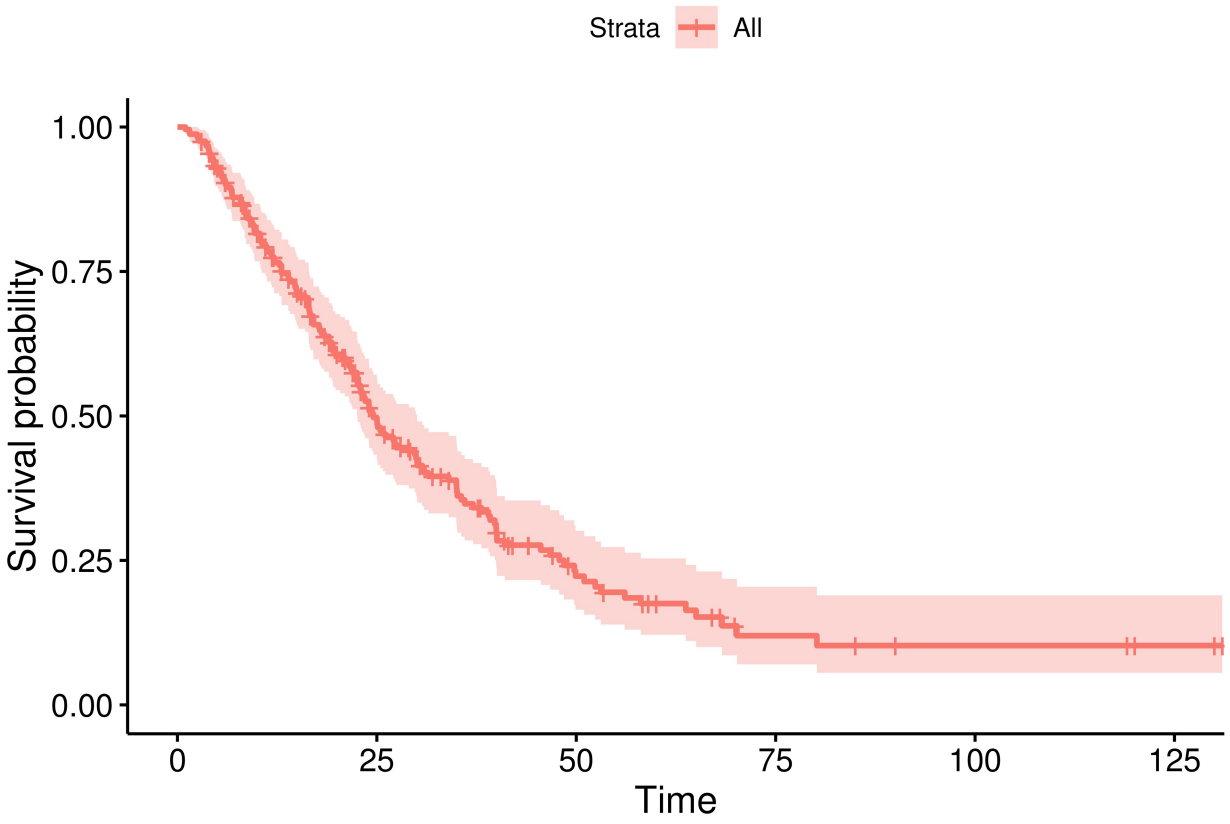
```
library(survival)
fit.surv <- survfit(Surv(time, status) ~ 1, data= df)

plot(fit.surv, xlab = "Months",
    ylab = "Estimated Probability of Survival")

library(survminer)
```

```
## Warning: package 'survminer' was built under R version 4.2.2

## Loading required package: ggpubr
```

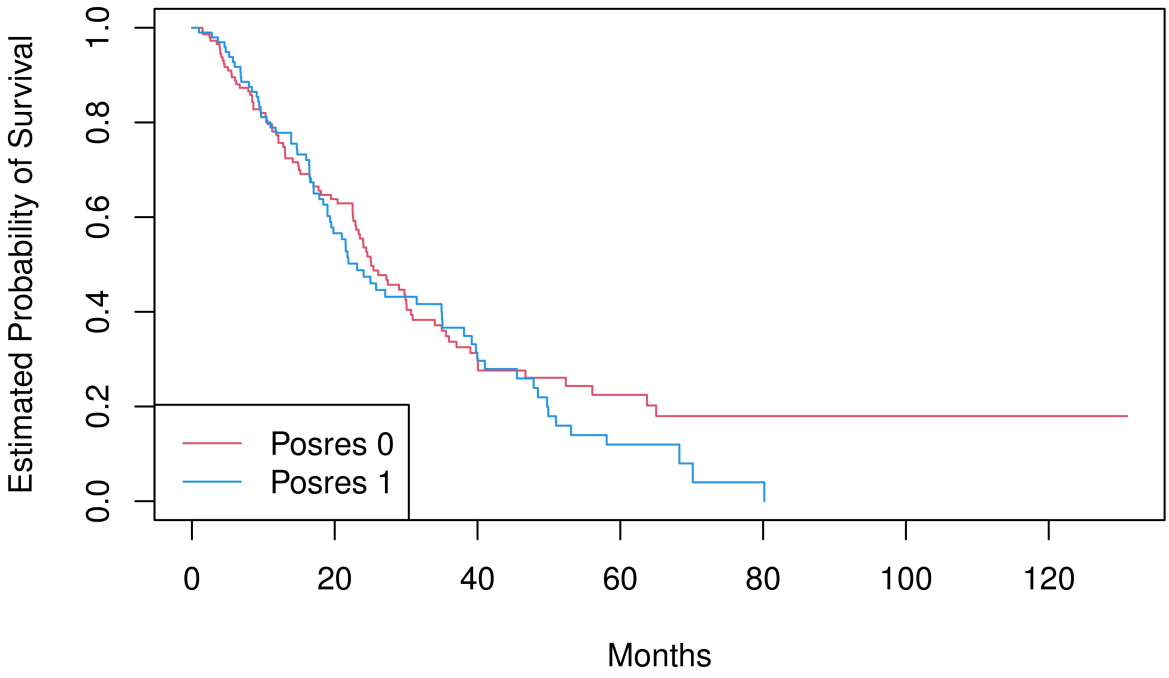
```
##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma
```

```
ggsurvplot(fit = fit.surv)
```

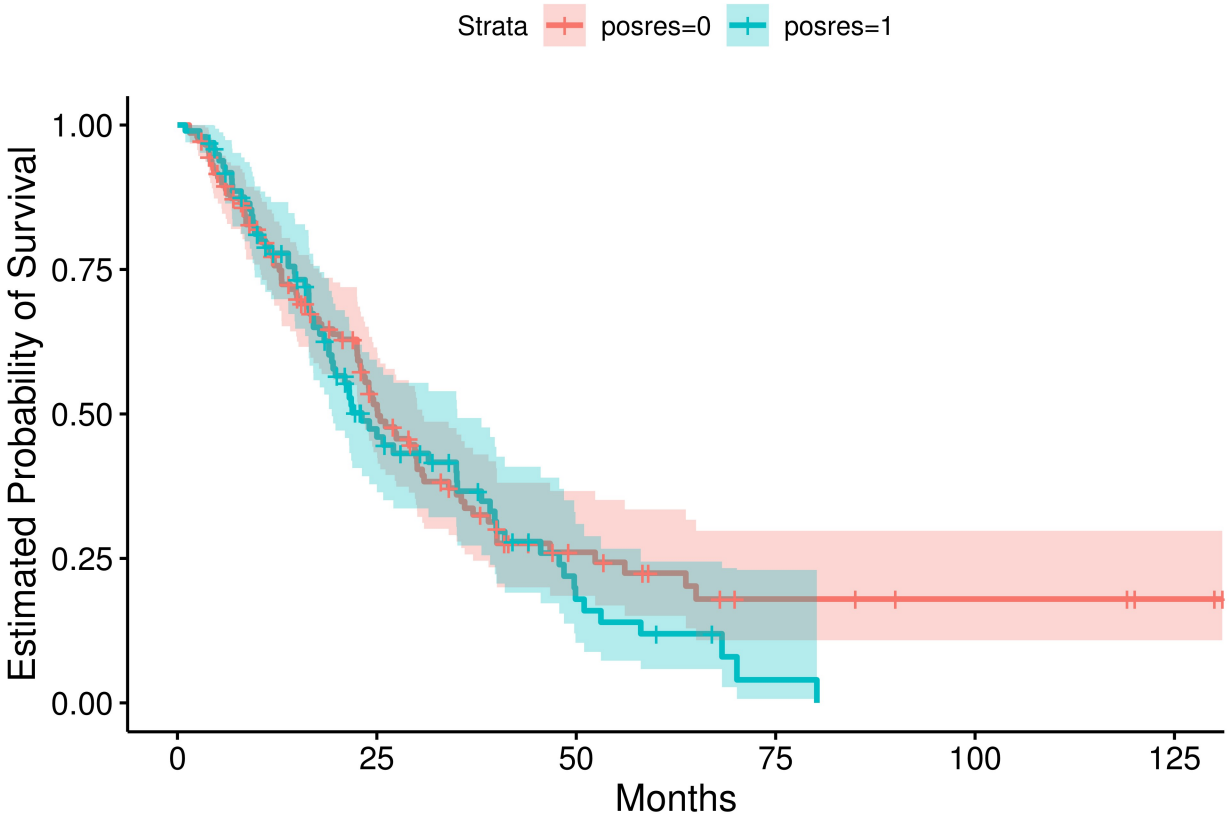





```
fit.surv <- survfit(Surv(time, status) ~ posres, data=df)
plot(fit.surv, xlab = "Months",
    ylab = "Estimated Probability of Survival", col = c(2,4))
legend("bottomleft",
  legend = c("Posres 0", "Posres 1"),
  col = c(2,4), lty = 1
 )
```




```
ggsurvplot(fit.surv,
           conf.int =T,
           xlab = "Months",
    ylab = "Estimated Probability of Survival")
```




```
logrank.test <- survdiff(Surv(time, status) ~ posres, data = df)
logrank.test
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ posres, data = df)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## posres=0 146       87     92.6     0.341     0.844
## posres=1  98       69     63.4     0.498     0.844
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

Our P value is 0.4, which is significantly greater than 0.05 indicating no evidence of a difference in survival between the two Posres groups. Note: This is valid till 50 months but after that there is a slight decrease in the survival probability of group 1, the probability of survival for group 0 remains constant.