HW 8

Nitish Neelagiri

November 22, 2015

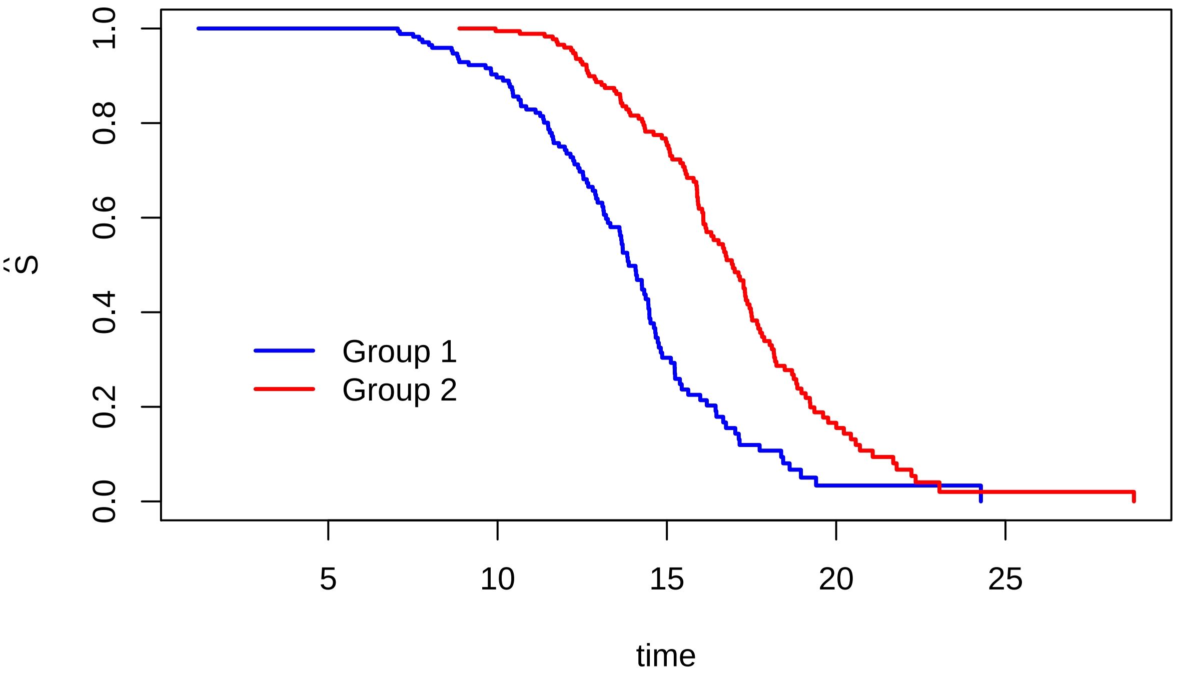
Question 1

(a)

```
setwd("~/")
library(survival)
survivalData <- read.csv("surv_times_data.csv")
attach(survivalData)

fit1_km = survfit(Surv(Y1, Delta1) ~ 1)
fit2_km = survfit(Surv(Y2, Delta2) ~ 1)

plot(fit1_km$time, fit1_km$surv, xlab = "time", ylab = expression(hat(S)),
  xlim = range(c(Y1, Y2)), ylim = c(0, 1), type = "s", col = "blue", lwd = 2)
lines(fit2_km$time, fit2_km$surv, type = "s", col = "red", lwd = 2)
legend(2, 0.4, legend = c("Group 1", "Group 2"), lwd = c(2, 2), col = c("blue", "red"),
  bty = "n")
```

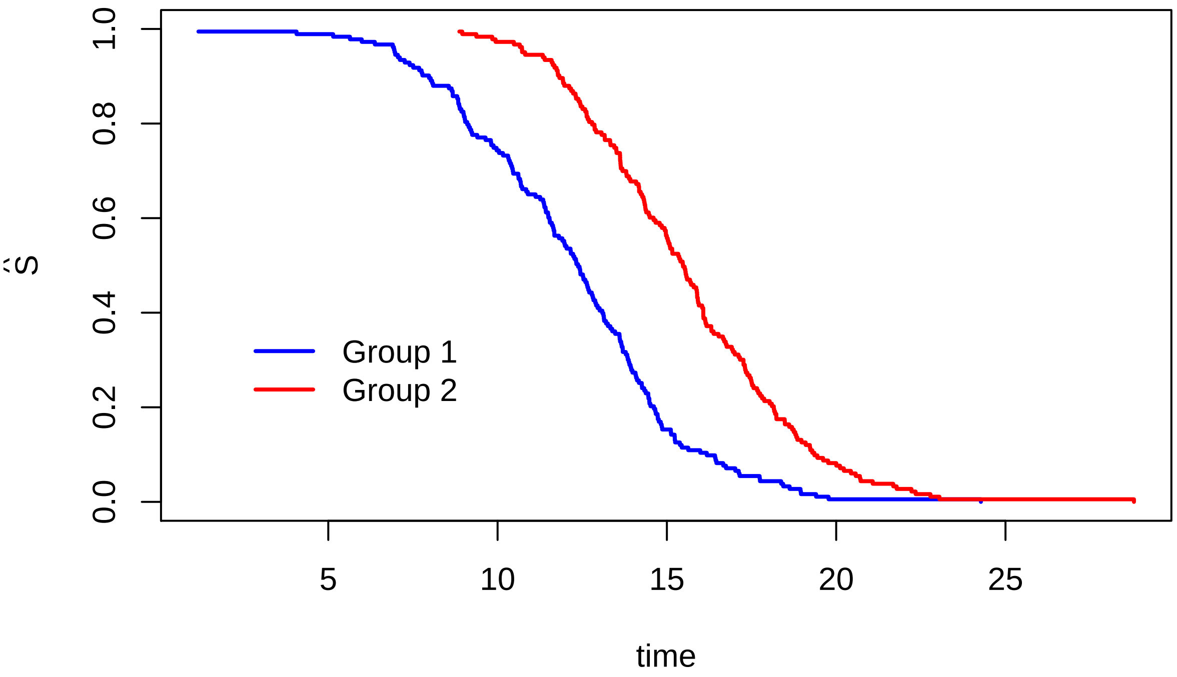


From the output plot, we can see that the survival probability for Group 2 is higher than Group 1 at a particular time event. So, Group 2 has better survival prognosis.

(b)

```
fit1_km_emp = survfit(Surv(Y1, rep(1, each = length(Y1))) ~ 1)
fit2_km_emp = survfit(Surv(Y2, rep(1, each = length(Y1))) ~ 1)

plot(fit1_km_emp$time, fit1_km_emp$surv, xlab = "time", ylab = expression(hat(S)), xlim = range(c(Y1, Y
lines(fit2_km_emp$time, fit2_km_emp$surv, type = "s", col = "red", lwd = 2)
legend(2, 0.4, legend = c("Group 1", "Group 2"), lwd = c(2, 2), col = c("blue", "red"),
  bty = "n")
```



We can observe from the plots that the empirical curve has low survival probability compared to Kaplan-Meier curve at a particular time event. This means that the inclusion of the censored data has increased the survival probability. Thus the censored data is more than the actual time event, meaning the right censored data.

(c)

```
time_hat_Y1 = c(0, fit1_km$time)
surv_hat_Y1 = c(1, fit1_km$surv)

time_hat_Y2 = c(0, fit2_km$time)
surv_hat_Y2 = c(1, fit2_km$surv)

mu_hat_km_Y1 = 0
mu_hat_km_Y2 = 0

for(i in 2:length(surv_hat_Y1)) {
  mu_hat_km_Y1 = mu_hat_km_Y1 + surv_hat_Y1[i - 1] * (time_hat_Y1[i] - time_hat_Y1[i - 1])
}
```

```
for(i in 2:length(surv_hat_Y2)) {
  mu_hat_km_Y2 = mu_hat_km_Y2 + surv_hat_Y2[i - 1] * (time_hat_Y2[i] - time_hat_Y2[i - 1])
}

#Mean Survival Time for Group 1
sum((time_hat_Y1[2:101] - time_hat_Y1[1:100]) * surv_hat_Y1[1:100])
```

```
## [1] 11.96973
```

```
#Mean Survival Time for Group 2
sum((time_hat_Y2[2:101] - time_hat_Y2[1:100]) * surv_hat_Y2[1:100])
```

```
## [1] 15.10546
```

(d)

```
#Quantile for Group 1:
quantile(fit1_km, c(0.25, 0.5 ,0.75))
```

```
## $quantile
##       25       50       75 
## 11.98686 13.87340 15.38351 
## 
## $lower
##       25       50       75 
## 11.34816 13.60025 14.76136 
## 
## $upper
##       25       50       75 
## 12.67745 14.48226 16.74518
```

```
#Quantile for Group 2:
quantile(fit2_km, c(0.25,0.5,0.75))
```

```
## $quantile
##       25       50       75 
## 15.04833 16.95023 18.82140 
## 
## $lower
##       25       50       75 
## 14.26997 16.16761 18.10029 
## 
## $upper
##       25       50       75 
## 15.86906 17.48058 20.00283
```

(e)

```
Y <- c(Y1, Y2)
Delta <- c(Delta1, Delta2)
GRP = factor(rep(0:1, each = 183))
survdiff(Surv(Y, Delta) ~ GRP)
```

```
## Call:
## survdiff(formula = Surv(Y, Delta) ~ GRP)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## GRP=0 183      110     62.8      35.5      51.9
## GRP=1 183      119    166.2      13.4      51.9
##
##  Chisq= 51.9  on 1 degrees of freedom, p= 5.78e-13
```

The chi squared value is 51.9 and p-value is 5.78e-13.

(f)

```
## General numerical optimization.
log_lik_Y = function(theta, Y, Delta) {
  mu = theta[1]
  sg = theta[2]

  obj_Y = sum(Delta * log(dnorm(Y, mu, sg))) +
    sum((1 - Delta) * log(1 - pnorm(Y, mu, sg)))

  return(obj_Y)
}

neg_log_lik_Y = function(theta, Y, Delta) { return(-log_lik_Y(theta, Y, Delta)) }

theta_0_Y1 = c(mean(Y1), sqrt(var(Y1)))
theta_0_Y2 = c(mean(Y2), sqrt(var(Y2)))

theta_hat_Y1 = optim(theta_0_Y1, neg_log_lik_Y, Y = Y1, Delta = Delta1)
theta_hat_Y2 = optim(theta_0_Y2, neg_log_lik_Y, Y = Y2, Delta = Delta2)

#MLEs of Group1:
#Mean :
theta_hat_Y1$par[1]
```

```
## [1] 13.92421
```

```
#SD:
theta_hat_Y1$par[2]
```

```
## [1] 3.101121
```

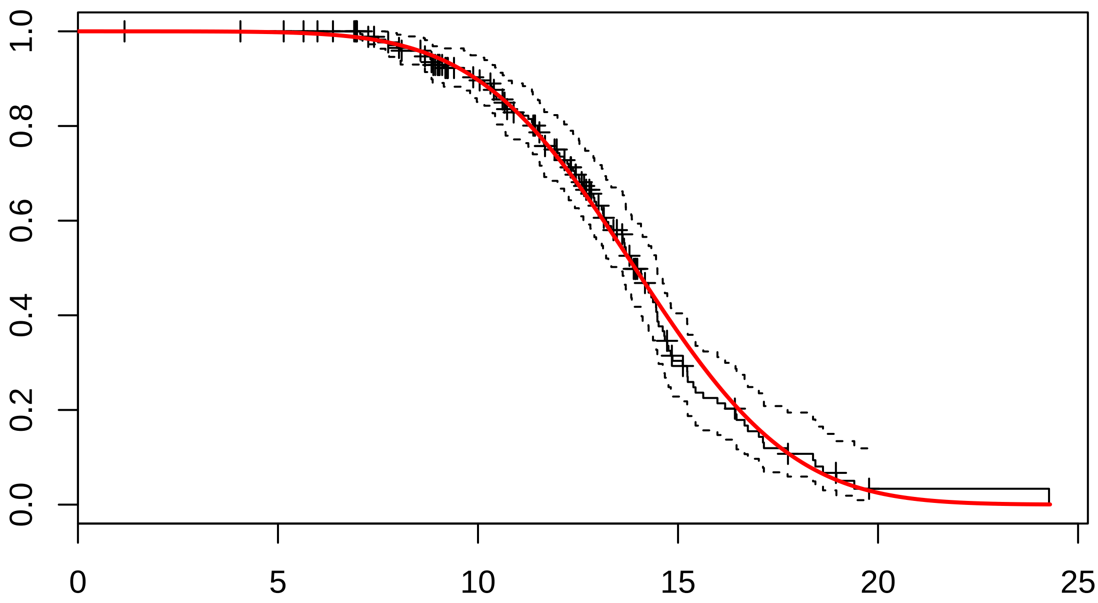
```
#MLEs of Group2:
#Mean:
theta_hat_Y2$par[1]
```

```
## [1] 17.00265
```

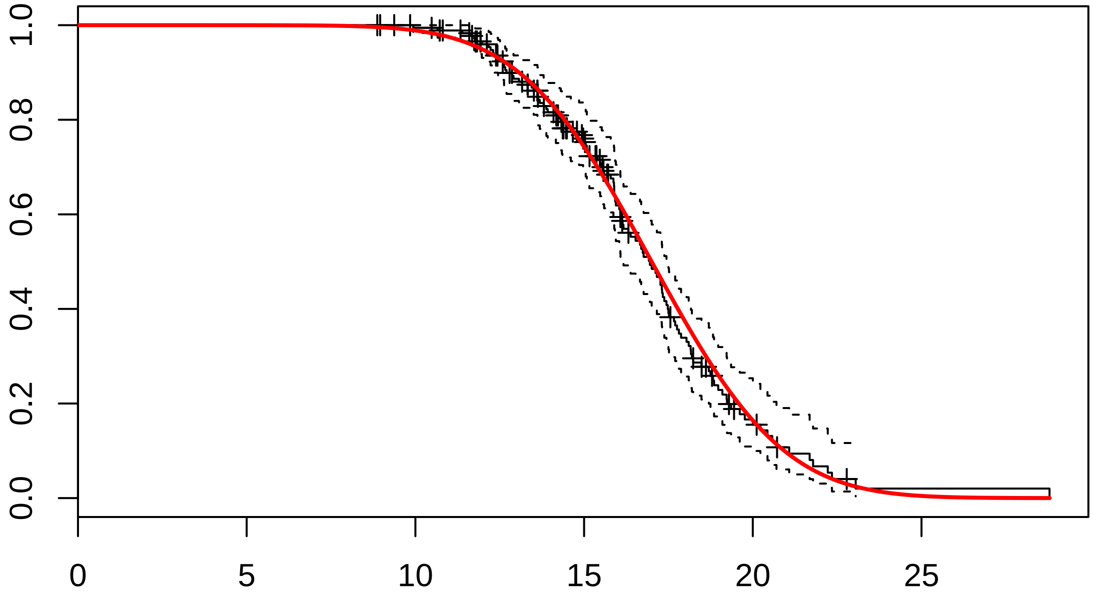
```
#SD:
theta_hat_Y2$par[2]
```

```
## [1] 3.06806
```

```
#Comparision of plots of Kaplan Meier and MLE based Survival curves. Red line represents MLE based surv
#Group 1
tm_Y1 = seq(0, 24.3, length = 1000)
S_hat_Y1 = 1 - pnorm(tm_Y1, theta_hat_Y1$par[1], theta_hat_Y1$par[2])
plot(fit1_km)
lines(tm_Y1, S_hat_Y1, lwd = 2, col = "red")
```



```
#Group 2
tm_Y2 = seq(0, 28.8, length = 1000)
S_hat_Y2 = 1 - pnorm(tm_Y2, theta_hat_Y2$par[1], theta_hat_Y2$par[2])
plot(fit2_km)
lines(tm_Y2, S_hat_Y2, lwd = 2, col = "red")
```



(g)

```
B <- 500
n <- nrow(survivalData)
delta_hat <- rep(NA,B)
for(b in 1:B){
  cat(".")
ii_1 <- sample(1:n, replace = TRUE)
ii_2 <- sample(1:n, replace = TRUE)
dta_b <- data.frame("Y1" = Y1[ii_1], "Y2" = Y2[ii_2], "Delta1" = Delta1[ii_1], "Delta2" = Delta2[ii_2])

theta_0_Y1_b <- with(dta_b, c(mean(Y1), sd(Y1)))
theta_0_Y2_b <- with(dta_b, c(mean(Y2), sd(Y2)))
fit_Y1_b <- with(dta_b, optim(theta_0_Y1_b, neg_log_lik_Y, Y = Y1, Delta = Delta1))
fit_Y2_b <- with(dta_b, optim(theta_0_Y2_b, neg_log_lik_Y, Y = Y2, Delta = Delta2))

delta_hat[b] = fit_Y1_b$par[1] - fit_Y2_b$par[1]
}
```

```
## ....................................................................................................
```

```
#95% Confidence Interval
quantile(delta_hat, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -3.875942 -2.325731
```