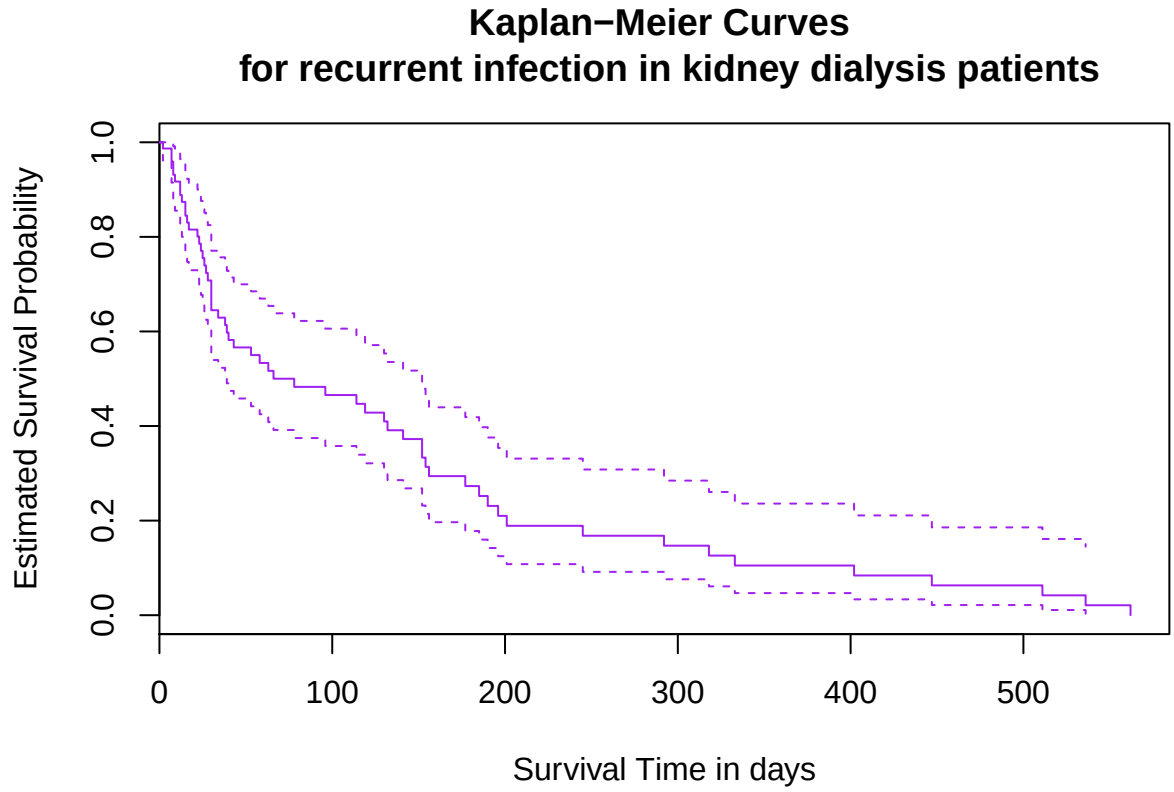
PSTAT175_lab_C

Mujie Wang

2019/10/28

Problem 1

```
#a)
library(survival)
data(kidney)
kidney.fit <- survfit(Surv(kidney$time,kidney$status) ~ 1)
plot(kidney.fit, main= "Kaplan-Meier Curves \n for recurrent infection in kidney dialysis patients",
     xlab = "Survival Time in days",
     ylab = "Estimated Survival Probability",
     conf.int = TRUE,col="purple"
     )
```



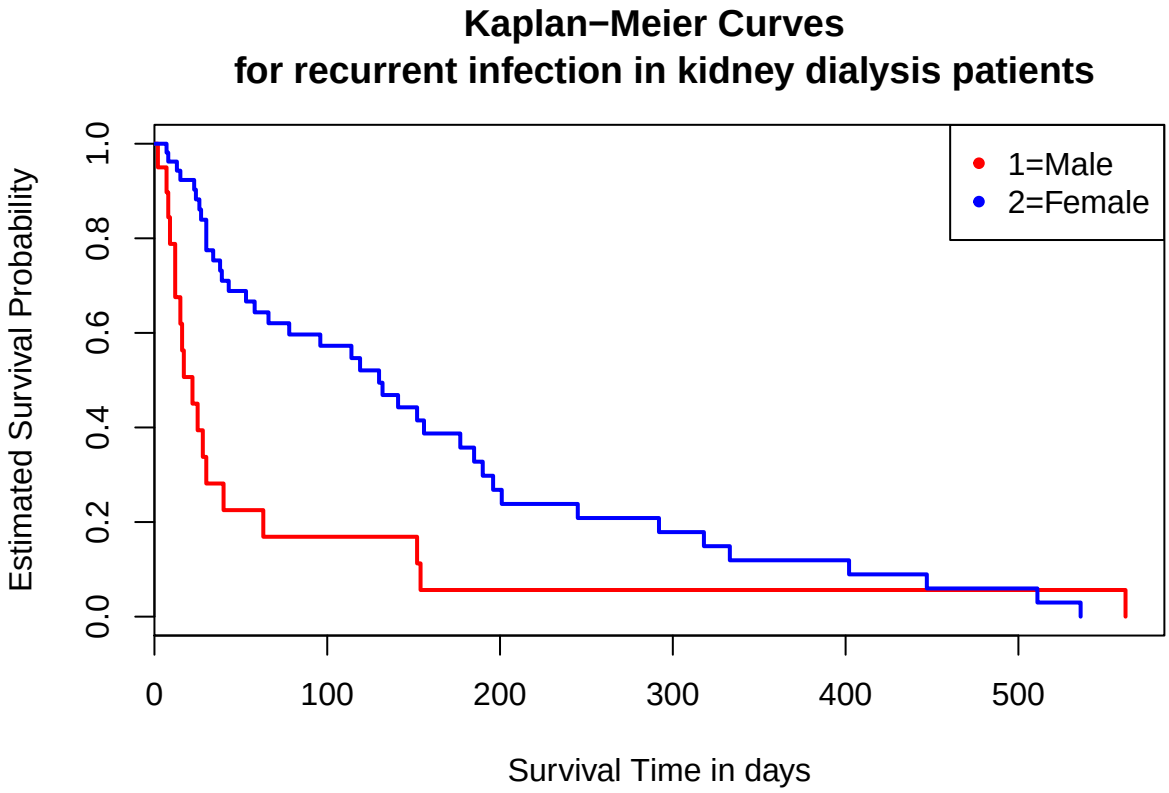

```
#b)
kidney.ft<- survdiff(Surv(kidney$time, kidney$status) ~ kidney$sex)
kidney.ft
```

```
## Call:
## survdiff(formula = Surv(kidney$time, kidney$status) ~ kidney$sex)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## kidney$sex=1 20       18     10.2      5.99      8.31
## kidney$sex=2 56       40     47.8      1.28      8.31
##
```

```
##  Chisq= 8.3  on 1 degrees of freedom, p= 0.00395
```

From this test, the p value is 0.004, which means there is statistically significant difference between the different sex groups survival rates.

```
#c)
sex <- as.factor(kidney$sex)
kidney.fit.sex <- survfit(Surv(kidney$time,kidney$status) ~ sex)
plot(kidney.fit.sex, main= "Kaplan-Meier Curves \n for recurrent infection in kidney dialysis patients"
     col=c(2,4),lwd=2,conf.int= FALSE, xlab = "Survival Time in days", ylab = "Estimated Survival Probal
legend("topright",legend=c("1=Male","2=Female"),
       col=c("red","blue"),pch=20)
```



From the plot, women have longer time until a recurrent infection before first 440 days and after that, men have longer time until a recurrent infection. By looking through the graph, it is clear that the women's curve are above men'a curve in the beginning, however, at about (440,500)days,women and men's curves meet, and after 500 days, women's curve is below men's curve. Since a higher survival function means a longer time until failure or death, we can say women have better suivival rates and longer time until a recurrent infection in the beginning and men have better survival rates after 500 days.

```
#d)
kidney.model <- coxph(Surv(kidney$time,kidney$status)~kidney$sex)
kidney.model
```

```
## Call:
## coxph(formula = Surv(kidney$time, kidney$status) ~ kidney$sex)
##
##              coef exp(coef) se(coef)     z      p
## kidney$sex -0.838     0.433    0.297 -2.82 0.0047
##
## Likelihood ratio test=7.07  on 1 df, p=0.00785
```

```
## n= 76, number of events= 58
```

```
exp(confint(kidney.model,level=0.95))
```

```
##                2.5 %    97.5 %
## kidney$sex 0.241936 0.7738447
```

The coefficient for sex is -0.8377 with hazard ratio is 0.4327. The hazard rate for men is 0.433 times of female patients, so that female have longer survival rates than male. The 95% confident interval is [0.241936,0.7738447] which include 0.433.

```
#e)
male_gp <- kidney[kidney$sex==1,]
(male_gp.fit <- survfit(Surv(time,status)~1,data=male_gp))
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = male_gp)
##
##       n  events  median 0.95LCL 0.95UCL
##      20      18      22      12      63
```

```
boxplot.stats(male_gp$time)$out #outerliner
```

```
## [1] 154 152 562
```

```
(male_42 <- kidney[42,])#data in row 42
```

```
##    id time status age sex disease frail
## 42 21  562      1  47   1     PKD   0.2
```

```
male_row <- (which(kidney$sex==1))
male_row_42 <- male_row[male_row!=42]
male_gp_42 <- kidney[male_row_42,]
(male_gp_42.fit <- survfit(Surv(time,status)~1,data=male_gp_42))
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = male_gp_42)
##
##       n  events  median 0.95LCL 0.95UCL
##      19      17      17      12      40
```

```
kidney_42 <- kidney[kidney$time!=562,]
(kidney_42.fit<- coxph(Surv(time,status)~sex, data = kidney_42))
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex, data = kidney_42)
##
##       coef exp(coef) se(coef)     z       p
## sex -1.488     0.226    0.319 -4.66 3.2e-06
##
## Likelihood ratio test=18.5  on 1 df, p=1.67e-05
## n= 75, number of events= 57
```

the observation in row 42 is concerned in the Kaplan???Meier estimate for the male group because the observed time is 562 in row 42 which is the outliner, the observation jump for 154 to 562 which is larger than other times for man patients.After removed the observation in row 42, the hazard ratio will drop to 0.2259 and make the survivor function of the two sexs more similiar.

problem 2

```
#a)
library(survival)
```

```
data(mgus)
mgus1 <- coxph(Surv(futime,death)~ sex, data=mgus)
mgus1
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ sex, data = mgus)
##
##          coef exp(coef) se(coef)    z     p
## sexmale 0.339     1.403    0.136 2.49 0.013
##
## Likelihood ratio test=6.28  on 1 df, p=0.0122
## n= 241, number of events= 225
```

Likelihood ratio test H0: Sm(t) = Sf(t); Ha: Sm(t)!= Sf(t),(m=male,f=female) The p-value is 0.01224 which is less than 0.05, we reject H0 which means the difference between sexs is significant.

```
#b)
#include 'sex'
cox <- coxph(Surv(futime,death)~age+alb+creat+hgb+mspike+sex, data=mgus)
cox
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + alb + creat + hgb +
##     mspike + sex, data = mgus)
##
##              coef exp(coef) se(coef)     z       p
## age      0.07035   1.07288  0.00855  8.22 2.2e-16
## alb     -0.25845   0.77225  0.20597 -1.25  0.2096
## creat    0.40527   1.49970  0.14710  2.76  0.0059
## hgb     -0.10683   0.89868  0.06058 -1.76  0.0778
## mspike   0.01063   1.01069  0.19907  0.05  0.9574
## sexmale  0.20552   1.22816  0.16502  1.25  0.2130
##
## Likelihood ratio test=97.2  on 6 df, p=0
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

```
#not include 'sex'
cox2 <- coxph(Surv(futime,death)~age+alb+creat+hgb+mspike, data=mgus)
cox2
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + alb + creat + hgb +
##     mspike, data = mgus)
##
##            coef exp(coef) se(coef)     z      p
## age     0.07108   1.07367  0.00853  8.33 <2e-16
## alb    -0.24266   0.78454  0.20391 -1.19 0.2340
## creat   0.42424   1.52843  0.13922  3.05 0.0023
## hgb    -0.09184   0.91225  0.05967 -1.54 0.1237
## mspike -0.01130   0.98877  0.19944 -0.06 0.9548
##
## Likelihood ratio test=95.6  on 5 df, p=0
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

```
#The first one is for null model where there is no covariate. The second one is what we need.
cox$loglik
```

```
## [1] -716.2474 -667.6617
```

```
#Compute the Likelihood Ratio
lrt <- 2*(cox$loglik[2]-cox2$loglik[2])
lrt
```

```
## [1] 1.562232
```

```
#Approximate p with the Chi-squared distribution
pchisq(lrt,df=1,lower.tail=FALSE)
```

```
## [1] 0.2113388
```

the p-value is 0.2113388 which is still less than 0.05, so the difference between sexs is significant.

c) The answers to part (a) and (b) are different because part (b) includes the five covariates age, alb, creat,hgb, and mspike; in part (a), the cox proportional hazard model does not keep the covariates constant, these will affect the final result.

```
#d)
cox3 <- coxph(Surv(futime,death)~ age+creat+hgb, data=mgus)
cox3
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + creat + hgb, data = mgus)
##
##           coef exp(coef) se(coef)     z       p
## age    0.07510   1.07799  0.00806  9.32 < 2e-16
## creat  0.44642   1.56270  0.13141  3.40 0.00068
## hgb   -0.12927   0.87874  0.04963 -2.60 0.00920
##
## Likelihood ratio test=112  on 3 df, p=0
## n= 198, number of events= 184
##    (43 observations deleted due to missingness)
```

```
cox4 <- coxph(Surv(futime,death)~ age+creat, data=mgus)
cox4
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + creat, data = mgus)
##
##          coef exp(coef) se(coef)    z       p
## age   0.07521   1.07811  0.00813 9.25 < 2e-16
## creat 0.48920   1.63101  0.13231 3.70 0.00022
##
## Likelihood ratio test=106  on 2 df, p=0
## n= 198, number of events= 184
##    (43 observations deleted due to missingness)
```

After removing the covariates hgb, the p-value decreases, which means the age of patients and the creatine level at MGUS diagnosis is significant in Hypothesis testing.

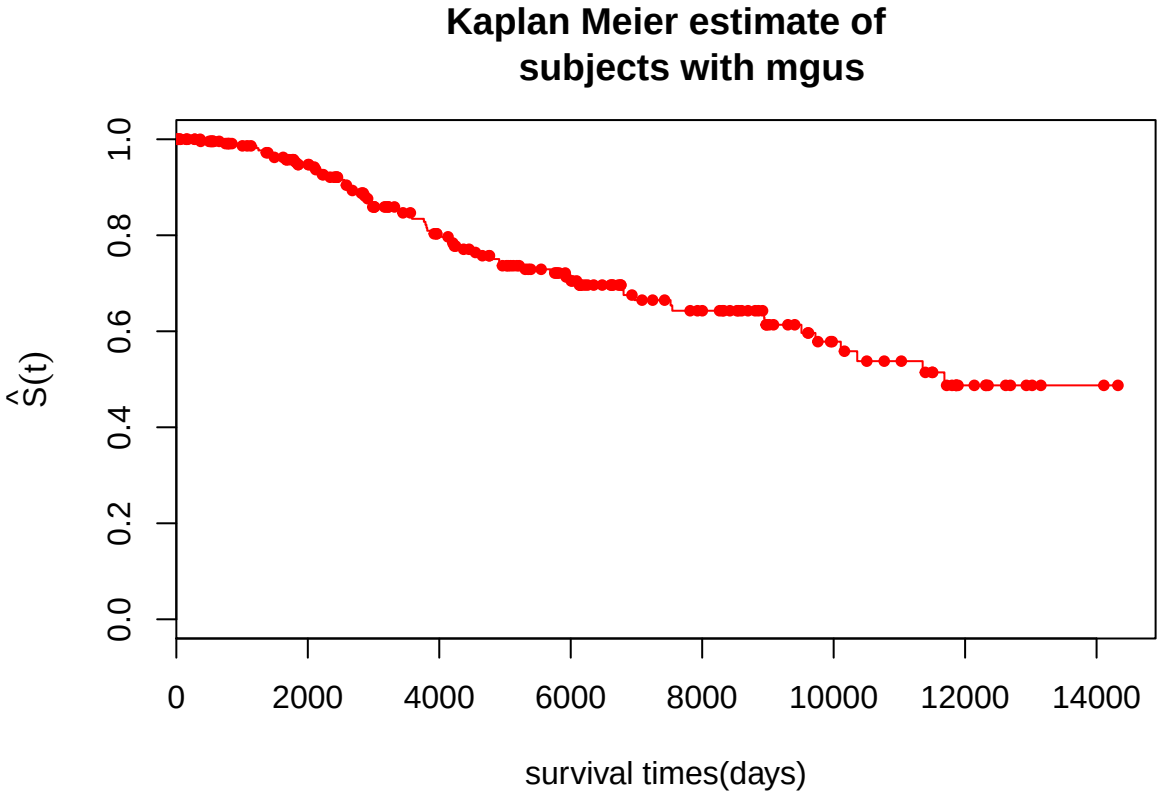
Problem 3

```
#a)
mgus2 <- mgus
mgus2["time"] <- 0
```

```r
mgus2["status"] <- 0

for (i in 1:dim(mgus2)[1]) {
  if(is.na(mgus2$pctime[i])){
    mgus2$time[i] <- mgus2$futime[i]
    mgus2$status[i] <- 0
    #indicate censord obs
  }
  else{
    mgus2$time[i] <- mgus2$pctime[i]
    mgus2$status[i] <- 1
    #non-censord obs
  }
}

mgus3.ft <- survfit(Surv(time, status)~1, data = mgus2)
par(mar=c(5,5,4,2))
plot(mgus3.ft, main = "Kaplan Meier estimate of
     subjects with mgus",
     xlab = "survival times(days)",
     ylab = expression(hat(S)(t)),
     col = "red", mark.time = TRUE, mark = 20, conf.int = FALSE)
```




```r
#b)
(mgus3b <- coxph(Surv(time,status)~mspike, data = mgus2))
```

```
## Call:
## coxph(formula = Surv(time, status) ~ mspike, data = mgus2)
```

```
##
##          coef exp(coef) se(coef)    z    p
## mspike -0.472     0.624    0.314 -1.5 0.13
##
## Likelihood ratio test=2.3  on 1 df, p=0.129
## n= 241, number of events= 64
```

The p-value is 0.129 which is greater than 0.05, the size of the monoclonal protein spike at MGUS diagnosis does not have an significant effect.

```
#c)
mgus3c_1 <- coxph(Surv(time,status)~sex+age+alb+creat+hgb+mspike, data=mgus2)
mgus3c_1
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + age + alb + creat +
##     hgb + mspike, data = mgus2)
##
##             coef exp(coef) se(coef)     z    p
## sexmale -0.30082   0.74021  0.33082 -0.91 0.36
## age      0.00194   1.00194  0.01417  0.14 0.89
## alb      0.20727   1.23031  0.36654  0.57 0.57
## creat   -0.28740   0.75021  0.63421 -0.45 0.65
## hgb     -0.09998   0.90485  0.11457 -0.87 0.38
## mspike  -0.64090   0.52682  0.39019 -1.64 0.10
##
## Likelihood ratio test=5.1  on 6 df, p=0.531
## n= 176, number of events= 46
##    (65 observations deleted due to missingness)
```

```
mgus3c_2 <- coxph(Surv(time,status)~sex+age+alb+creat+hgb, data=mgus2)
mgus3c_2
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + age + alb + creat +
##     hgb, data = mgus2)
##
##             coef exp(coef) se(coef)     z    p
## sexmale -0.26821   0.76475  0.33067 -0.81 0.42
## age      0.00216   1.00216  0.01411  0.15 0.88
## alb      0.08760   1.09155  0.36472  0.24 0.81
## creat   -0.36166   0.69652  0.63430 -0.57 0.57
## hgb     -0.07695   0.92593  0.11311 -0.68 0.50
##
## Likelihood ratio test=2.29  on 5 df, p=0.807
## n= 176, number of events= 46
##    (65 observations deleted due to missingness)
```

```
#Compute the Likelihood Ratio
lrt <- 2*(mgus3c_1$loglik[2]-mgus3c_2$loglik[2])
lrt
```

```
## [1] 2.808573
```

```
#Approximate p with the Chi-squared distribution
pchisq(lrt,df=1,lower.tail=FALSE)
```

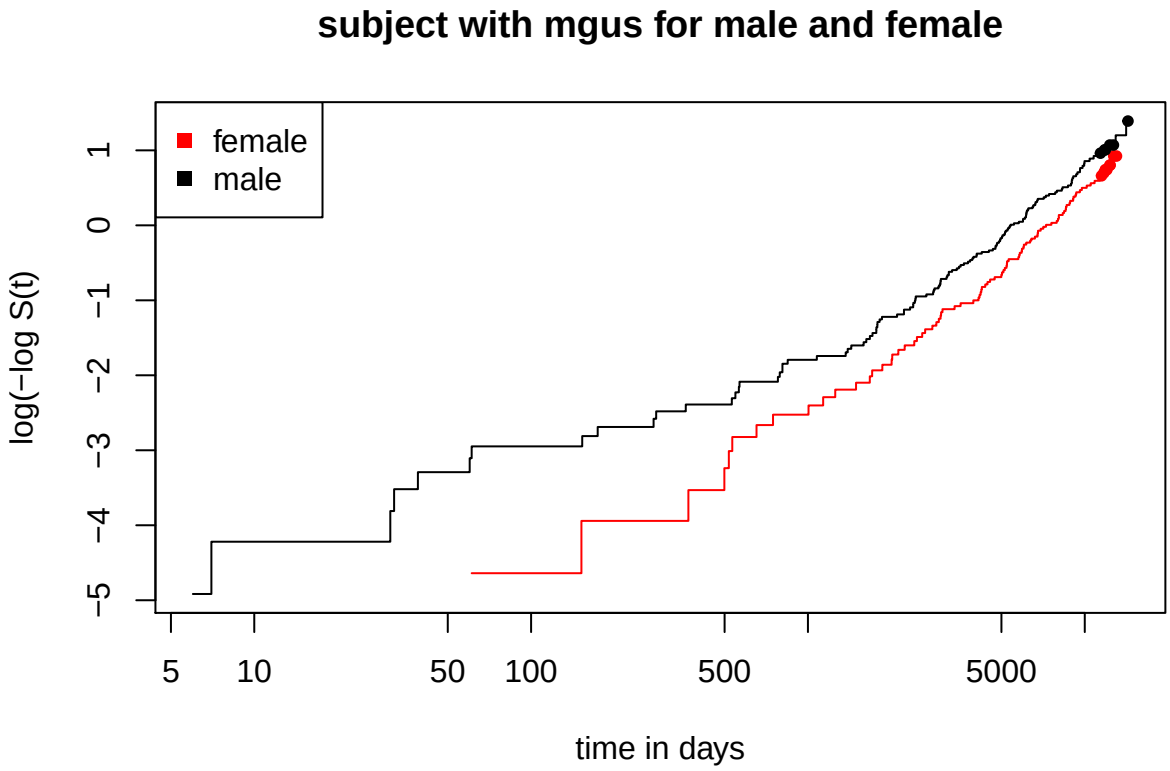
```
## [1] 0.09376174
```

The p-value is 0.09376174 which is greater than 0.05. The size of the monoclonal protein spike at MGUS diagnosis does not have a significant effect on time until a further disease is present.

Problem 4

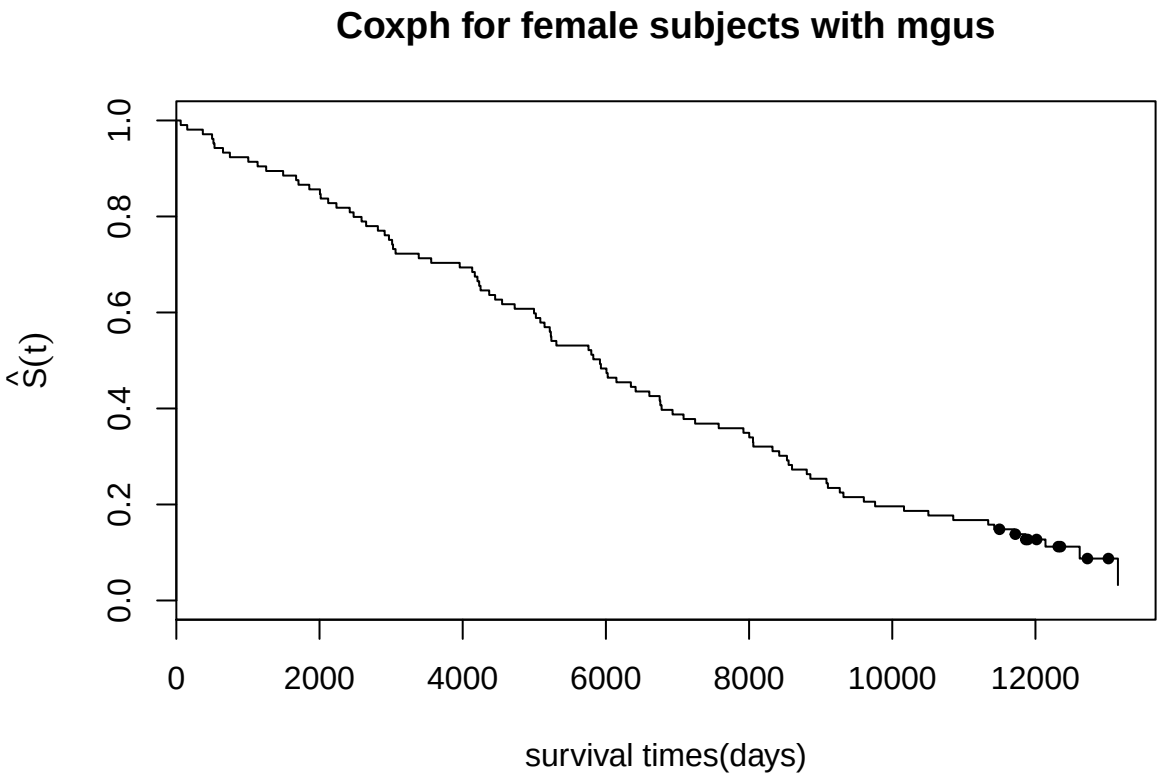
```
#4a
mgus4a.ft <- survfit(Surv(futime,death)~sex, data = mgus)
plot(mgus4a.ft, fun = "cloglog",
  main = "subject with mgus for male and female",
  xlab = "time in days", ylab = "log(-log S(t)",
  col=2:1, mark.time=TRUE, mark = 20)

legend("topleft", c("female","male"),pch = 15, col = 2:1)
```

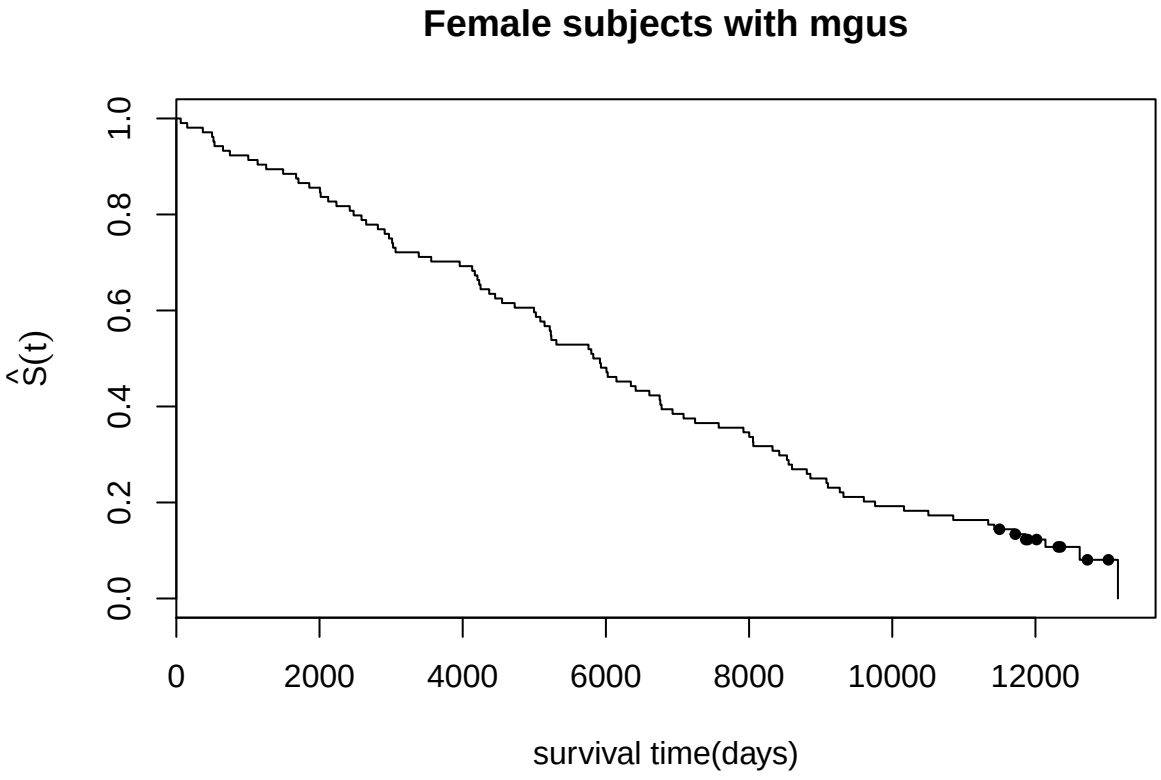


There is no evidence that proportional hazards model is not appropriate. it is because the lines are likely parallel to each other and no cross appears.

```
#4b
mgus_female <- mgus[mgus$sex=="female",]
mgus_female_0 <- coxph(Surv(futime,death)~1, data = mgus_female)
par(mar=c(5,5,4,2))
plot(survfit(mgus_female_0),
     main = "Coxph for female subjects with mgus",
     xlab = "survival times(days)", ylab = expression(hat(S)(t)),
     mark.time = T, mark = 20, conf.int = FALSE)
```




```
par(mar=c(5,5,4,2))
mgus_female_fit <- survfit(Surv(futime,death)~1, data = mgus_female)
plot(mgus_female_fit, main = "Female subjects with mgus",
     xlab = "survival time(days)", ylab = expression(hat(S)(t)),
     mark.time = T, mark = 20, conf.int = FALSE)
```



the plot seems very similar. that means the cox proportion hazard model gives a resonable fit about the data of female.

```
#4c
cox.zph(mgus1, global = FALSE)
```

```
##             rho chisq     p
## sexmale -0.0833  1.53 0.216
```

The p-value is 0.216 which is greater than 0.05. Therefore, the model is not significantly divergent from the proportional hazards model. As a result, we are justified in using the proportional hazards assumption in our modeling of the effect of sexs.