labc_zhongyun zhang

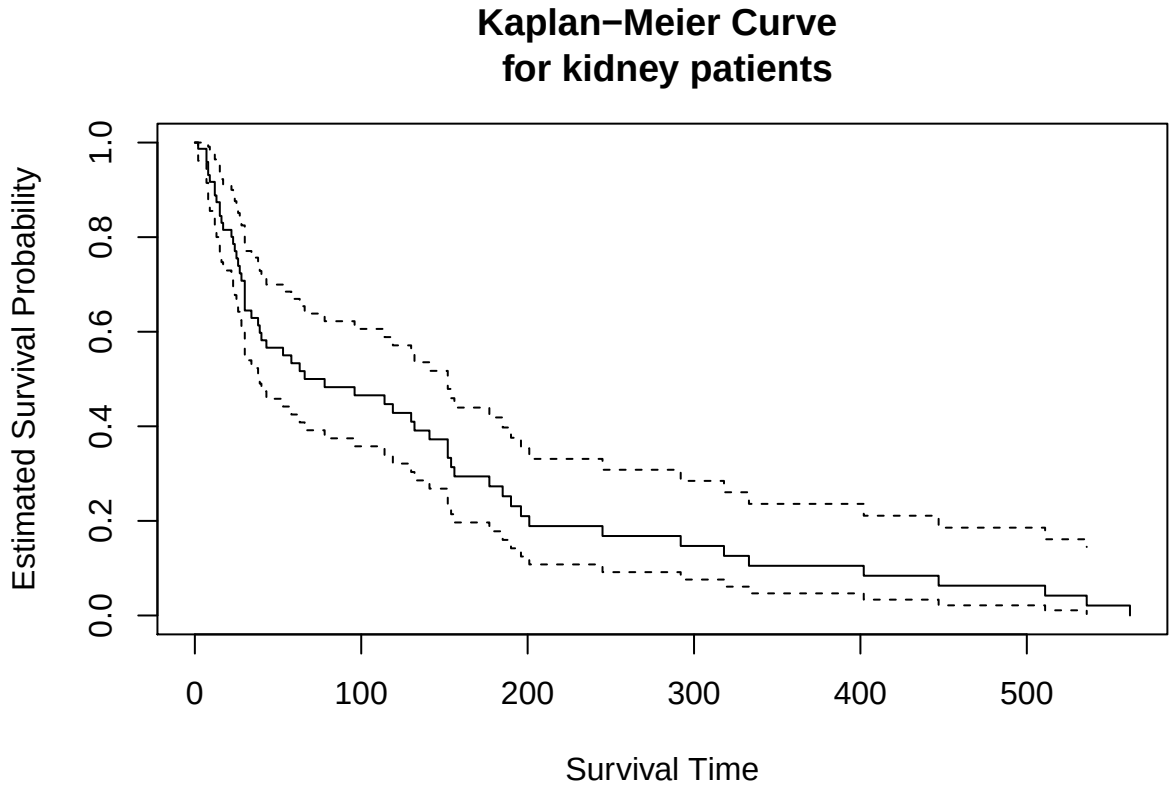
zhongyun zhang

2019/10/28

Problem 1

a.Plot the Kaplan–Meier estimate of the survivor function.

```
#install.packages("survival")
library(survival)
data(kidney)
kidney.surv <- Surv(kidney$time,kidney$status)
kidney.fit = survfit(Surv(kidney$time,kidney$status)~1,data=kidney)
plot(kidney.fit,xlab="Survival Time",ylab="Estimated Survival Probability",
    main="Kaplan-Meier Curve \n for kidney patients" )
```



(b) Use survdiff to perform a logrank test on whether or not there is a difference between the sexes. The kidney$sex variable is coded as 1 for male subjects and 2 for female subjects. What do you conclude from this test?

```
survdiff(Surv(time,status)~sex, data=kidney)
```

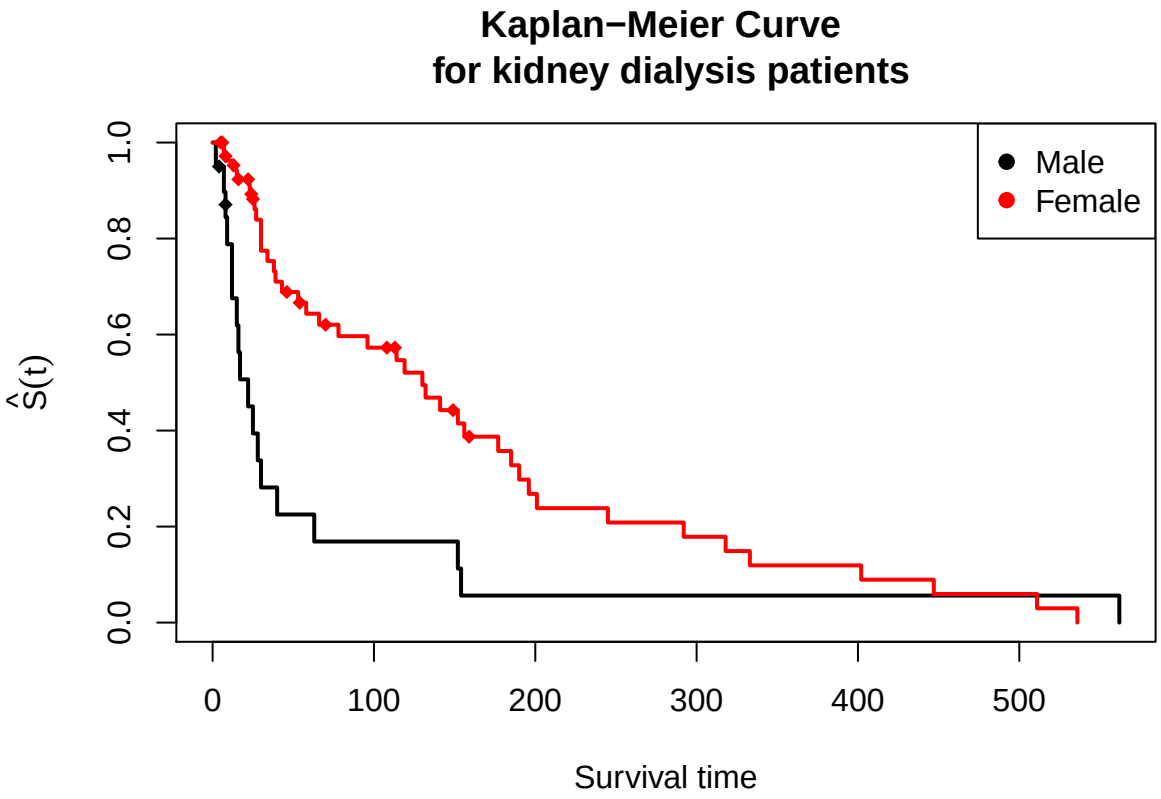
```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = kidney)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 20       18     10.2      5.99      8.31
```

```
## sex=2 56       40     47.8      1.28      8.31
##
##  Chisq= 8.3  on 1 degrees of freedom, p= 0.004
```

The p-value is 0.004<0.05, we conclude that there is statistically significant difference between the men and women's groups survival rates.

(c) Create a plot that compares the Kaplan–Meier estimates of the survivor functions for the two sexes separately. Describe what you see in this plot to confirm your test result from part (b).

```
sexkm = survfit(kidney.surv~sex,data=kidney)
par(mar=c(5,5,4,2))
plot(sexkm,xlab="Survival time",ylab = expression(hat(S)(t)),lwd=2, col=1:2,
     mark.time = TRUE,mark=18,main="Kaplan-Meier Curve \n for kidney dialysis patients")
legend("topright",legend=c("Male","Female"),col=1:2,pch=rep(19,2))
```



Generally speaking, the estimate for the women' survial rate is a lot higher. A higher survival function means a longer time until failure or death. The curves converge near the end of the survival time, around days 440, which means that at that time, men and women may have nearly similar survival rates. This confirms that there is statistically significant differences between the treatment groups survival rates.

(d) Use the coxph function to estimate the hazard proportion between the two sexes. Explain to someone who is not familiar with the model how you would interpret the meaning of this number. For instance, what does it say about the different probabilities of going a month without an infection? Give a 95% confidence interval for your parameter estimate.

```
#Use compare the two treatment groups (without considering covariates).
cox0<-coxph(Surv(kidney$time,kidney$status)~sex, data = kidney)
cox0
```

```
## Call:
## coxph(formula = Surv(kidney$time, kidney$status) ~ sex, data = kidney)
```

```
##
##        coef exp(coef) se(coef)      z       p
## sex -0.8377    0.4327   0.2966 -2.824 0.00474
##
## Likelihood ratio test=7.07  on 1 df, p=0.007848
## n= 76, number of events= 58
```

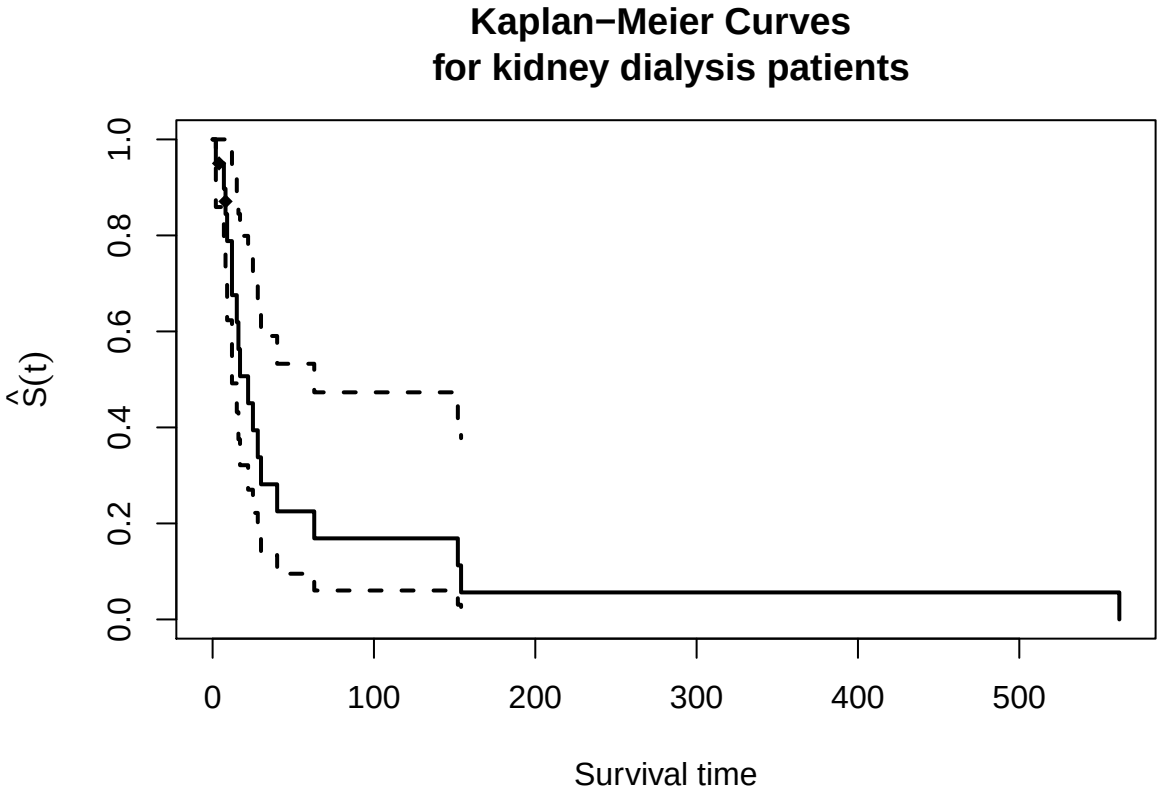
```
(exp(confint(cox0,level = 0.95)))
```

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

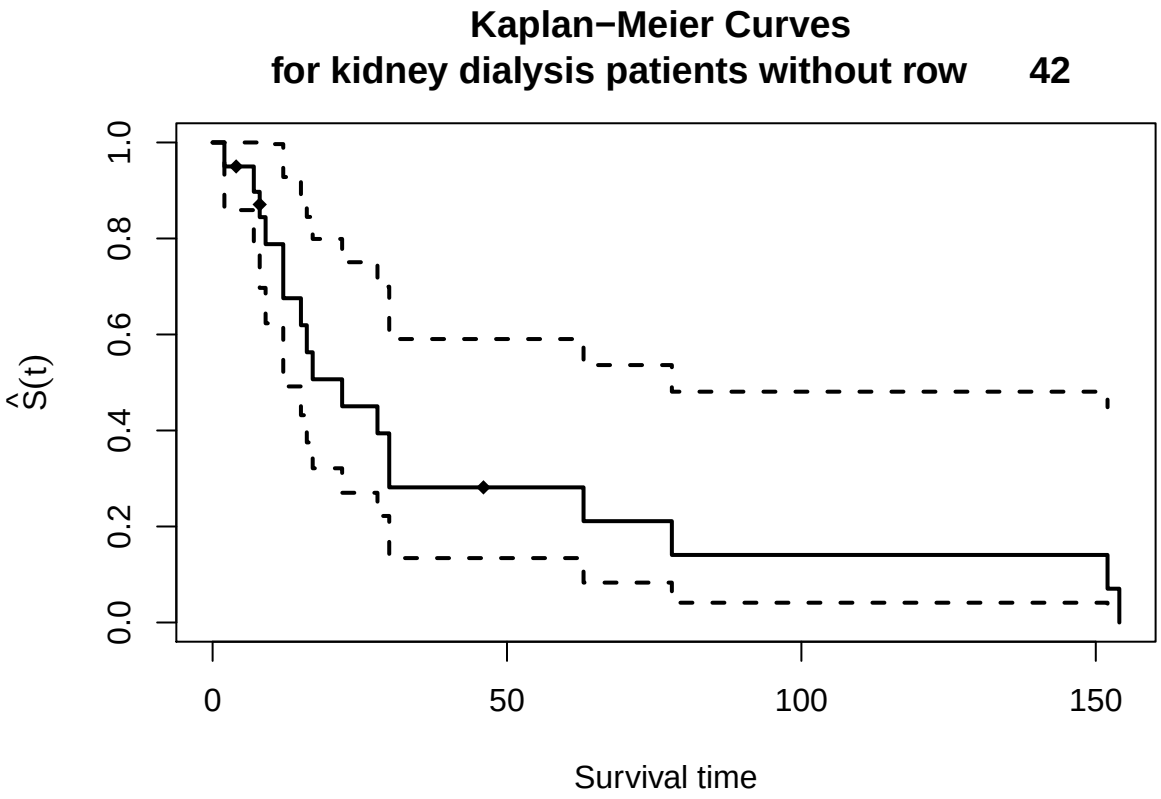
The estimated hazard proportion between the two sexes is 0.4327. The confidence interval of the harzard ratio is (0.241936, 0.7738447) Female increase the narzard by a factor of 43.27%, which means female decrease the risk of getting infection by 56.73%. Women's average probability of live without infection is 56.73% higher than men.

(e) Looking at the Kaplan–Meier estimate for the male group, the observation in row 42 is concerning to me. Why am I concerned? Explore how your analysis would change if that one observation was removed.

```
kidney.male=survfit(Surv(time,status)~1,subset = (kidney$sex==1),data=kidney)
par(mar=c(5,5,4,2))
plot(kidney.male,main="Kaplan-Meier Curves \n for kidney dialysis patients",
     xlab="Survival time", ylab=expression(hat(S)(t)),lwd=2,
     mark.time = TRUE,mark=18)
```




```
kidney.42<-kidney[-c(42),]
kidney.42fit<-survfit(Surv(time,status)~1,subset=(kidney.42$sex==1),data=kidney)
par(mar=c(5,5,4,2))
plot(kidney.42fit,main="Kaplan-Meier Curves \n for kidney dialysis patients without row     42",xlab="
```



The observation in row 42 may be outlied from Kaplan-Meier estimates and it is a censored data.It is improtant because it has the largest survival time. After removing the point, the range of x value decrease from more than 500 to about 150, the estimate survival rate increased

Problem 2

(a) Test the hypothesis that there is a significant difference between the two sexes. Report a P-value and a conclusion.

```
#install.packages("survival")
library(survival)
data("mgus")
mgus.futime<-mgus$futime
mgus.death<-mgus$death
mgus.surv<-Surv(mgus.futime,mgus.death)
cox2 <- coxph(Surv(futime,death)~sex,data=mgus)
cox2
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ sex, data = mgus)
##
##           coef exp(coef) se(coef)     z      p
## sexmale 0.3385    1.4029   0.1360 2.489 0.0128
##
## Likelihood ratio test=6.28  on 1 df, p=0.01224
## n= 241, number of events= 225
```

h0=there is no significant difference between two sexes h1=there is significant difference between two sexes p-value=0.01224<0.05 reject h0, we can conclude that there is significant difference between two cases.

(b) Re-run the test of the difference between the sexes, but use the covariates age, alb, creat, hgb, and

mspike to control for differences in the groups. Report a P-value and a conclusion.

```
cox1 <- coxph(Surv(futime,death)~age+alb+creat+hgb+mspike+sex,data=mgus)
cox1
```

```
## Call:
## coxph(formula = Surv(futime, death) ~ age + alb + creat + hgb +
##     mspike + sex, data = mgus)
##
##               coef exp(coef)  se(coef)      z       p
## age       0.070350  1.072884  0.008554  8.225 < 2e-16
## alb      -0.258449  0.772249  0.205974 -1.255 0.20957
## creat     0.405267  1.499702  0.147102  2.755 0.00587
## hgb      -0.106833  0.898676  0.060577 -1.764 0.07780
## mspike    0.010634  1.010691  0.199070  0.053 0.95740
## sexmale   0.205519  1.228162  0.165020  1.245 0.21298
##
## Likelihood ratio test=97.17  on 6 df, p=< 2.2e-16
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

h0= controling variables, there is no significant difference between two sexes h1= controling variables, there is significant difference between two sexes p-value =0.21298>0.05 we fail to reject h0. We can conclude that there is no significant differences between two sex.

(c) How might we explain the difference in the answers to part (a) and (b)?

```
coxph(mgus.surv~mgus$age+mgus$alb+mgus$creat+mgus$mspike,data=mgus)
```

```
## Call:
## coxph(formula = mgus.surv ~ mgus$age + mgus$alb + mgus$creat +
##     mgus$mspike, data = mgus)
##
##                   coef exp(coef)  se(coef)      z       p
## mgus$age      0.070159  1.072678  0.008596  8.162 3.3e-16
## mgus$alb     -0.356407  0.700188  0.192765 -1.849 0.06447
## mgus$creat    0.449750  1.567921  0.140453  3.202 0.00136
## mgus$mspike   0.017836  1.017996  0.199295  0.089 0.92869
##
## Likelihood ratio test=93.26  on 4 df, p=< 2.2e-16
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

```
coxph(mgus.surv~mgus$age+mgus$alb+mgus$creat,data=mgus)
```

```
## Call:
## coxph(formula = mgus.surv ~ mgus$age + mgus$alb + mgus$creat,
##     data = mgus)
##
##                  coef exp(coef)  se(coef)      z       p
## mgus$age     0.070154  1.072673  0.008599  8.158 3.4e-16
## mgus$alb    -0.353168  0.702459  0.189401 -1.865 0.06223
## mgus$creat   0.450582  1.569225  0.139912  3.220 0.00128
##
## Likelihood ratio test=93.25  on 3 df, p=< 2.2e-16
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

P value of alb, hgb, msike is more than 0.05, so it should not include in this model, the new model with age and creat goe very small p p-vlue. So age and create should be a better fit for this model. The model in a only have sex in and excluded others. In this way sex is important for this model.However, when otehr variables are added into the modelm they hare more helpful than sex does.

(d) Propose a set of covariates that you think best fits the futime data in a not-too-complicated way. Justify your choices with regression results.

```
(coxph(mgus.surv~mgus$age+mgus$alb+mgus$creat+mgus$hgb+mgus$mspike, data=mgus))
```

```
## Call:
## coxph(formula = mgus.surv ~ mgus$age + mgus$alb + mgus$creat +
##     mgus$hgb + mgus$mspike, data = mgus)
##
##                   coef exp(coef)  se(coef)      z       p
## mgus$age     0.071085  1.073672  0.008532  8.332 < 2e-16
## mgus$alb    -0.242658  0.784540  0.203911 -1.190 0.23404
## mgus$creat   0.424241  1.528430  0.139218  3.047 0.00231
## mgus$hgb    -0.091841  0.912250  0.059666 -1.539 0.12374
## mgus$mspike -0.011297  0.988767  0.199441 -0.057 0.95483
##
## Likelihood ratio test=95.61  on 5 df, p=< 2.2e-16
## n= 176, number of events= 165
##    (65 observations deleted due to missingness)
```

```
(coxph(mgus.surv~mgus$age+mgus$creat,data=mgus))
```

```
## Call:
## coxph(formula = mgus.surv ~ mgus$age + mgus$creat, data = mgus)
##
##                coef exp(coef) se(coef)     z        p
## mgus$age   0.075207  1.078107 0.008129 9.251  < 2e-16
## mgus$creat 0.489196  1.631005 0.132307 3.697 0.000218
##
## Likelihood ratio test=105.8  on 2 df, p=< 2.2e-16
## n= 198, number of events= 184
##    (43 observations deleted due to missingness)
```

The set of covariates best fits data is age and creat. The p-value of alb,hgb and mspike are 0.23404,0.12374,0.95483 respectively. The p value are greater than 0.05, so alb, hgb and mspike should not be included in the model. We should only use age and creat.Create a model with age and creat, their p-value get smaller. Age and creat best fit.

Question 3

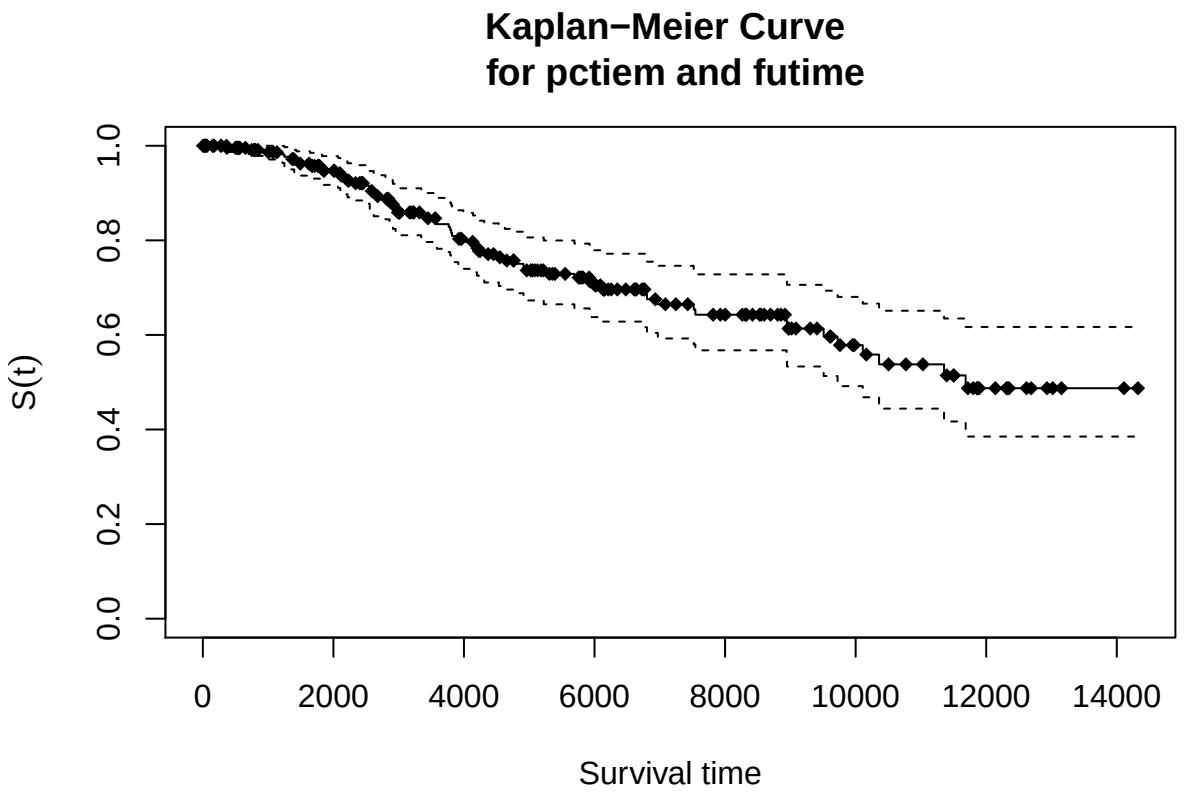
(a) The vector mgus$pctime contains the time after original diagnosis when the subject was diagnosed with a further, more sever
$uptime mgus$futime. Create a survival object from the pctime and futime measurements. Plot the KM estimate of the survivor function for this data.

```
head(mgus)
```

```
##   id age    sex dxyr pcdx pctime futime death alb creat  hgb mspike
## 1  1  78 female   68 <NA>     NA    748     1 2.8   1.2 11.5    2.0
## 2  2  73 female   66   LP   1310   6751     1  NA    NA   NA    1.3
## 3  3  87   male   68 <NA>     NA    277     1 2.2   1.1 11.2    1.3
## 4  4  86   male   69 <NA>     NA   1815     1 2.8   1.3 15.3    1.8
## 5  5  74 female   68 <NA>     NA   2587     1 3.0   0.8  9.8    1.4
```

```
## 6  6  81   male   68 <NA>     NA    563     1 2.9   0.9 11.5    1.8
```

```
mgus$newtime <-mgus$pctime
mgus$newtime[is.na(mgus$pctime)==TRUE]<-mgus$futime[is.na(mgus$pctime)==TRUE]
mgus$nevent[is.na(mgus$pctime)==TRUE]<-0
mgus$nevent[is.na(mgus$pctime)==FALSE]<-1
mgusnew.fit = survfit(Surv(mgus$newtime,mgus$nevent)~1,data=mgus)
plot(mgusnew.fit,main="Kaplan-Meier Curve \n for pctiem and futime",
     xlab="Survival time",ylab = expression(hat(S)(t)),mark.time=TRUE, mark=18)
```



(b) Use a Cox proportional model to see if mspike has an effect on the time until a further disease is present. Report a P-value and a conclusion.

```
cox3 <- coxph(Surv(newtime,nevent)~mspike,data=mgus)
cox3
```

```
## Call:
## coxph(formula = Surv(newtime, nevent) ~ mspike, data = mgus)
##
##           coef exp(coef) se(coef)      z     p
## mspike -0.4723    0.6236   0.3142 -1.503 0.133
##
## Likelihood ratio test=2.3  on 1 df, p=0.129
## n= 241, number of events= 64
```

h0= mspike has an effect onthe time until a further disease is present h1= mspike does not have an effect ont he time until a further disease is present. p-value =0.129>0.05 fail to reject h0. We can conclude that mspike has an effect onthe time until a further disease is present.

(c) Test the effect of mspike again but control for sex, age, alb, creat, and hgb. Report a P-value and a conclusion.

```
cox4 <- coxph(Surv(newtime,nevent)~mspike+sex+age+alb+creat+hgb,data=mgus)
cox4
```

```
## Call:
## coxph(formula = Surv(newtime, nevent) ~ mspike + sex + age +
##     alb + creat + hgb, data = mgus)
##
##              coef exp(coef)  se(coef)      z     p
## mspike  -0.640898  0.526819  0.390192 -1.643 0.100
## sexmale -0.300816  0.740214  0.330824 -0.909 0.363
## age      0.001943  1.001944  0.014168  0.137 0.891
## alb      0.207265  1.230309  0.366535  0.565 0.572
## creat   -0.287399  0.750213  0.634209 -0.453 0.650
## hgb     -0.099981  0.904855  0.114570 -0.873 0.383
##
## Likelihood ratio test=5.1  on 6 df, p=0.5307
## n= 176, number of events= 46
##    (65 observations deleted due to missingness)
```

h0= controlling other coviariates mspike has an effect onthe time until a further disease is present h1= controlling other coviariates, mspike does not have an effect ont he time until a further disease is present. p-value =0.100>0.05 fail to reject h0. We can conclude that mspike has an effect onthe time until a further disease is present.

Question 4

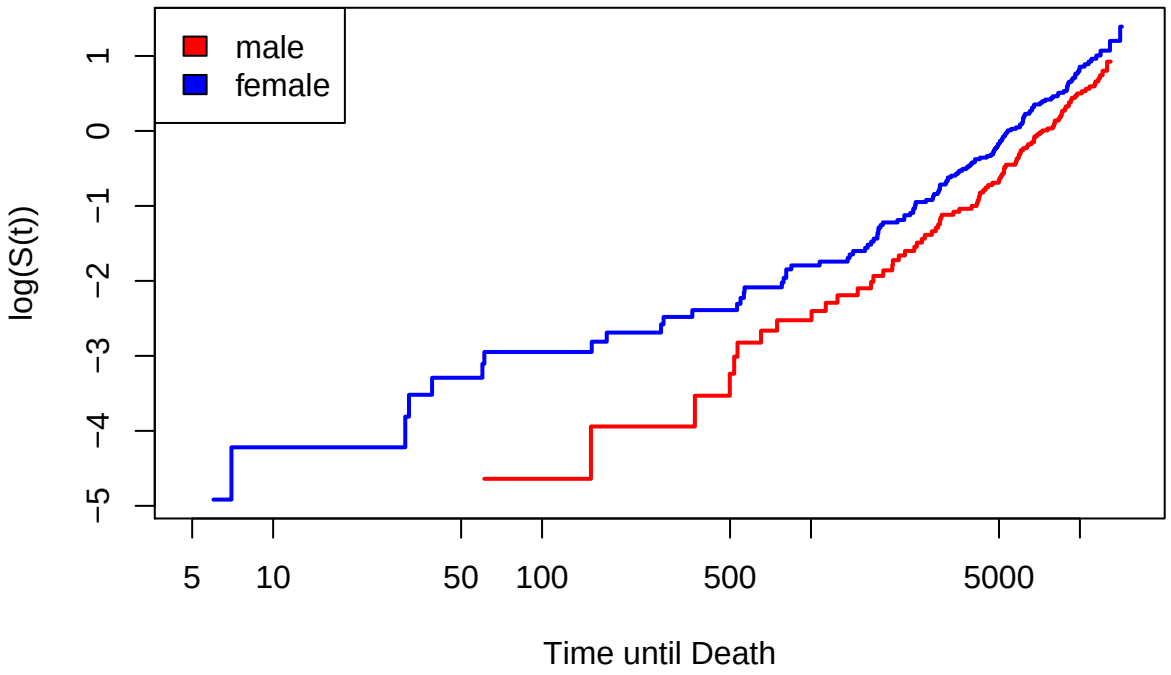
(a) Plot the -log-log graphs of the estimates of the survival functions for the men and women in the study. Is there evidence that the proportional hazards model is not appropriate?

```
mgus.surv <- Surv(mgus.futime,mgus.death)

mgus.fits<-survfit(Surv(mgus.futime,mgus.death)~sex,data=mgus)
plot(mgus.fits,lwd=2,col=c(2,4), xlim=c(5,15000),
fun="cloglog",xlab="Time until Death",ylab="log(S(t))")
```

```
## Warning in xy.coords(x, y, xlabel, ylabel, log): 1 x value <= 0 omitted
## from logarithmic plot
```

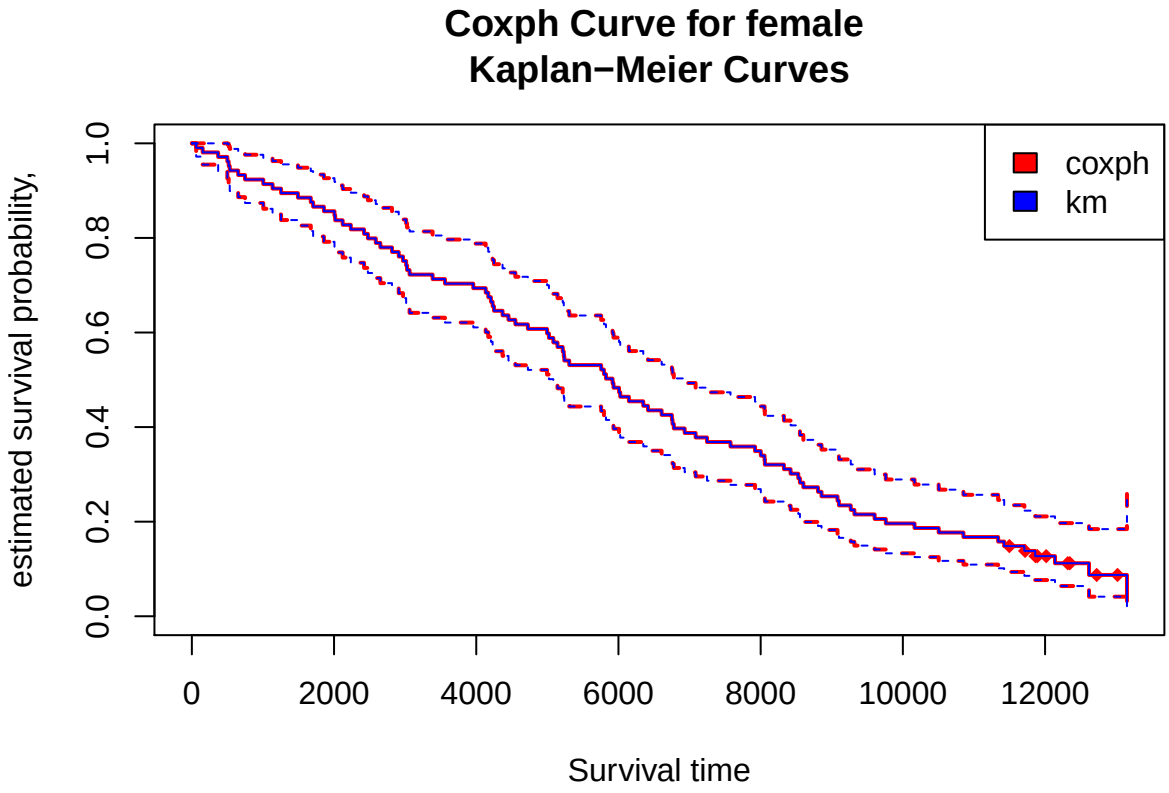
```
legend('topleft',c("male","female"),fill = c("red","blue"))
```



(b) Plot the estimated survival function from coxph for an average female subject, and then the KM estimate using only the women in the study. Compare the two estimates. Does it look like the model gives a reasonable fit?

```
mgus.female.cox<-survfit(coxph(Surv(mgus$futime,mgus$death)~mgus$sex,
                               subset=(mgus$sex=="female")))
#coxph curve for females
plot(mgus.female.cox,main="Coxph Curve for female \n Kaplan-Meier Curves ",
     xlab="Survival time", ylab="estimated survival probability,",lwd=2,
     mark.time = TRUE,mark=18,col='red')

#km curve
mgus.female.km<-survfit(coxph(Surv(mgus$futime,mgus$death)~1,subset=(mgus$sex=="female")))
lines(mgus.female.km,col='blue')
legend('topright',c("coxph","km"),fill=c("red","blue"))
```



Two plots are the same, the cox proportional hazard model gives a reasonable fit

(c) Use the cox.zph function to perform a test to see if the model is significantly divergent from the proportional hazards model. Interpret the result. Do you think that we are justified in using the proportional hazards assumption in our modeling of the effect of sex?

```
cox.zph(coxph(mgus.surv~mgus$sex,data=mgus))
```

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

h0: the coxph model is not significantly divergent ha: the model is significantly divergent from coxph The p-value =0.216>0.05, so we can conclude that we fail to reject H0. The model is not significantly hazard model. We are justified in using the proportional hazards assumption in modeling of the effect of sex.