STAD92 Project

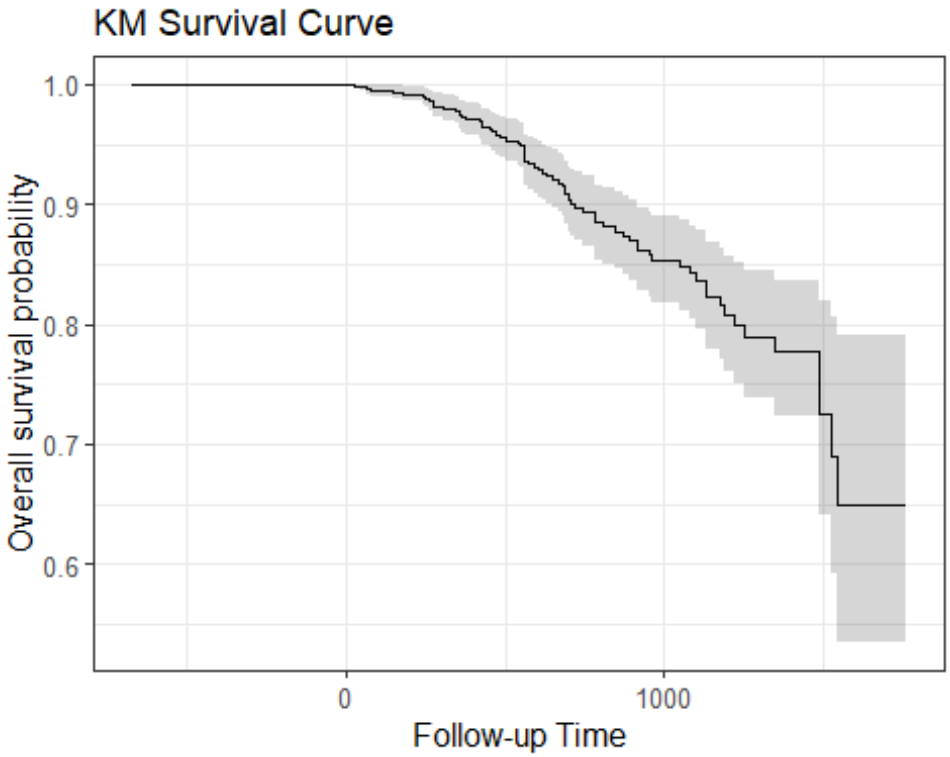
Congyu Hang

2023-08-27

Q1

a). Draw a KM survival curve for the entire data(without considering any independent variable). Also calculate median survival time along with 95% confidence interval.

```
library(tidyverse)
library(readxl)
pd<- read_excel("C:/Users/ASUS/Desktop/STAD92/STAD92_S2023_Project_Data.xlsx")
#convert status variable to binary form
pd = pd %>% mutate(status =
                     case_when(vstatus == "Alive" ~ 0,
                               TRUE ~ 1), .after = vstatus)
library(survival)
library(survminer)

library(ggsurvfit)
survfit2(Surv(f_time,status)~1, data=pd) %>%
    ggsurvfit() +
  labs(
    x = "Follow-up Time",
    y = "Overall survival probability",
    title = "KM Survival Curve"
  )+
  add_confidence_interval()+
  add_quantile(y_value = 0.5, color = "red", linewidth = 0.75)
```



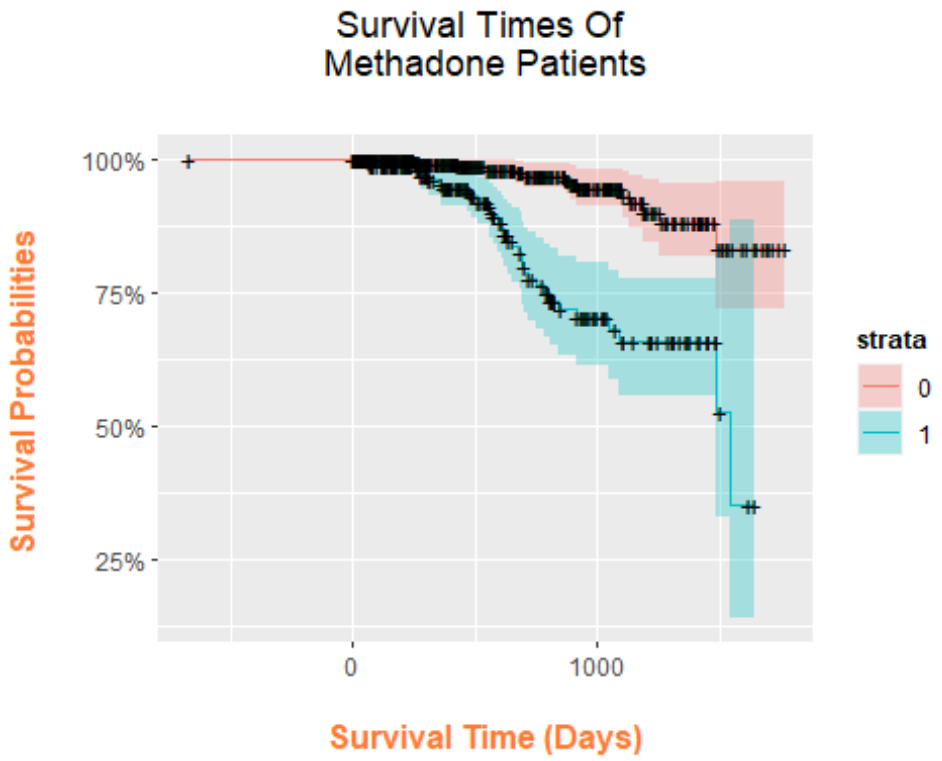

```
#get median survival time
(result.km <- survfit(Surv(f_time,status)~1, data=pd,conf.type = "log-log"))
```

```
## Call: survfit(formula = Surv(f_time, status) ~ 1, data = pd, conf.type = "log-log")
##
##         n events median 0.95LCL 0.95UCL
## [1,] 914     73     NA      NA      NA
```

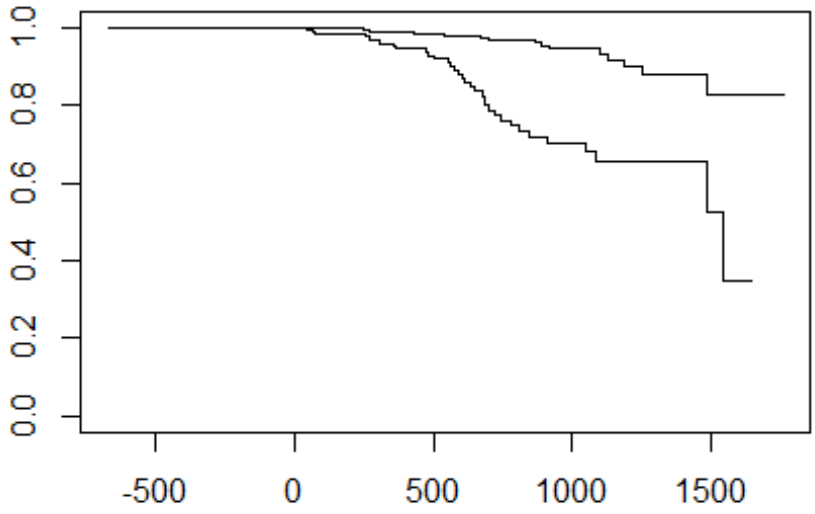
The output indicates that the median survival time is NA, according to the survival curve above, the median survival time is infinity.

b). Draw another KM curve that shows two survival curves (two curves in one plot): one curve for ulcer = No, and another curve for ulcer=Yes. Also carry out a log-rank test to compare the survival between these two categories of ulcer presence. Interpret your outputs

```
library(ggplot2)
library(ggfortify)

m1 <- survfit(Surv(f_time,status)~ulcer, data=pd)


autoplot(m1) +
 labs(x = "\n Survival Time (Days) ", y = "Survival Probabilities \n",
 title = "Survival Times Of \n Methadone Patients \n") +
 theme(plot.title = element_text(hjust = 0.5),
 axis.title.x = element_text(face="bold", colour="#FF7A33", size = 12),
 axis.title.y = element_text(face="bold", colour="#FF7A33", size = 12),
 legend.title = element_text(face="bold", size = 10))
```




```
plot(m1)
```



Log-rank test H0: Common survival curves for ulcer =No and ulcer=Yes

```
survdiff(Surv(f_time,status)~ulcer, data=pd)

## Call:
## survdiff(formula = Surv(f_time, status) ~ ulcer, data = pd)
##
```

```
## n=678
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## ulcer=0 452       15     33.8      10.4      33.9
## ulcer=1 226       34     15.2      23.2      33.9
##
##  Chisq= 33.9  on 1 degrees of freedom, p= 6e-09
```

Since p= 6e-09 < 0.05, reject H0 and conclude different ulcer generate different survival curves.

Q2

a) Fit a Cox-PH model using the given data. Use t_stge, ulcer and thick as independent variables. You can remove independent variables from the model if you find necessary. Interpret the outputs of your model

Cox PH Model

```
library(survival)
pdnew = na.omit(pd)
m2 <- coxph(
  Surv(f_time,status)~t_stage+ulcer+thick,data = pdnew)
summary(m2)
```

```
## Call:
## coxph(formula = Surv(f_time, status) ~ t_stage + ulcer + thick,
##     data = pdnew)
##
##   n= 670, number of events= 48
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## t_stageII  -1.82872   0.16062  1.42046 -1.287    0.198
## t_stageIII -0.43031   0.65031  1.07055 -0.402    0.688
## t_stageIV   0.28322   1.32740  1.05452  0.269    0.788
## t_stageV    0.60187   1.82553  1.25772  0.479    0.632
## ulcer       1.36322   3.90875  0.34306  3.974 7.07e-05 ***
## thick       0.01197   1.01204  0.08106  0.148    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## t_stageII     0.1606     6.2259  0.009924     2.600
## t_stageIII    0.6503     1.5377  0.079775     5.301
## t_stageIV     1.3274     0.7534  0.168032    10.486
## t_stageV      1.8255     0.5478  0.155172    21.476
## ulcer         3.9087     0.2558  1.995392     7.657
## thick         1.0120     0.9881  0.863381     1.186
##
## Concordance= 0.772  (se = 0.039 )
## Likelihood ratio test= 43.26  on 6 df,   p=1e-07
## Wald test            = 34.43  on 6 df,   p=6e-06
## Score (logrank) test = 45.76  on 6 df,   p=3e-08
```

```
extractAIC(m2)
```

```
## [1]   6.0000 481.3554
```

Since p-value of thick is large, remove thick from the model.

```
m3 <- coxph(
  Surv(f_time,status)~t_stage+ulcer,data = pdnew)
summary(m3)
```

```
## Call:
## coxph(formula = Surv(f_time, status) ~ t_stage + ulcer, data = pdnew)
##
##   n= 670, number of events= 48
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## t_stageII  -1.8278    0.1608   1.4205 -1.287    0.198
## t_stageIII -0.4188    0.6578   1.0677 -0.392    0.695
## t_stageIV   0.3072    1.3596   1.0418  0.295    0.768
## t_stageV    0.6860    1.9858   1.1176  0.614    0.539
## ulcer       1.3782    3.9678   0.3273  4.211 2.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## t_stageII     0.1608     6.2199  0.009933     2.602
## t_stageIII    0.6578     1.5201  0.081149     5.333
## t_stageIV     1.3596     0.7355  0.176455    10.476
## t_stageV      1.9858     0.5036  0.222156    17.751
## ulcer         3.9678     0.2520  2.088960     7.537
##
## Concordance= 0.759  (se = 0.036 )
## Likelihood ratio test= 43.24  on 5 df,   p=3e-08
## Wald test            = 34.34  on 5 df,   p=2e-06
## Score (logrank) test = 45.42  on 5 df,   p=1e-08
```

```
extractAIC(m3)
```

```
## [1]   5.0000 479.3771
```

Since p-value of t_stage is large, remove thick from the model

```
m4 <- coxph(
  Surv(f_time,status)~ulcer,data = pdnew)
summary(m4)
```

```
## Call:
## coxph(formula = Surv(f_time, status) ~ ulcer, data = pdnew)
##
##   n= 670, number of events= 48
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## ulcer 1.6459    5.1858   0.3126 5.265  1.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## ulcer     5.186     0.1928      2.81      9.57
##
## Concordance= 0.709  (se = 0.036 )
## Likelihood ratio test= 30.65  on 1 df,   p=3e-08
```

```
## Wald test            = 27.72  on 1 df,   p=1e-07
## Score (logrank) test = 34.38  on 1 df,   p=5e-09
```

```
extractAIC(m4)
```

```
## [1]   1.0000 483.9661
```

However, the AIC of m3 is the smallest, so we apply the Cox-PH model with t_stage and ulcer.

exp(coef)

t_stageII 0.1608

HR of t_stageII vs. t_stageI is 0.1608

t_stageIII 0.6578

HR of t_stageIII vs. t_stageI is 0.6578

t_stageIV 1.3596

HR of t_stageIV vs. t_stageI is 1.3596

t_stageV 1.9858

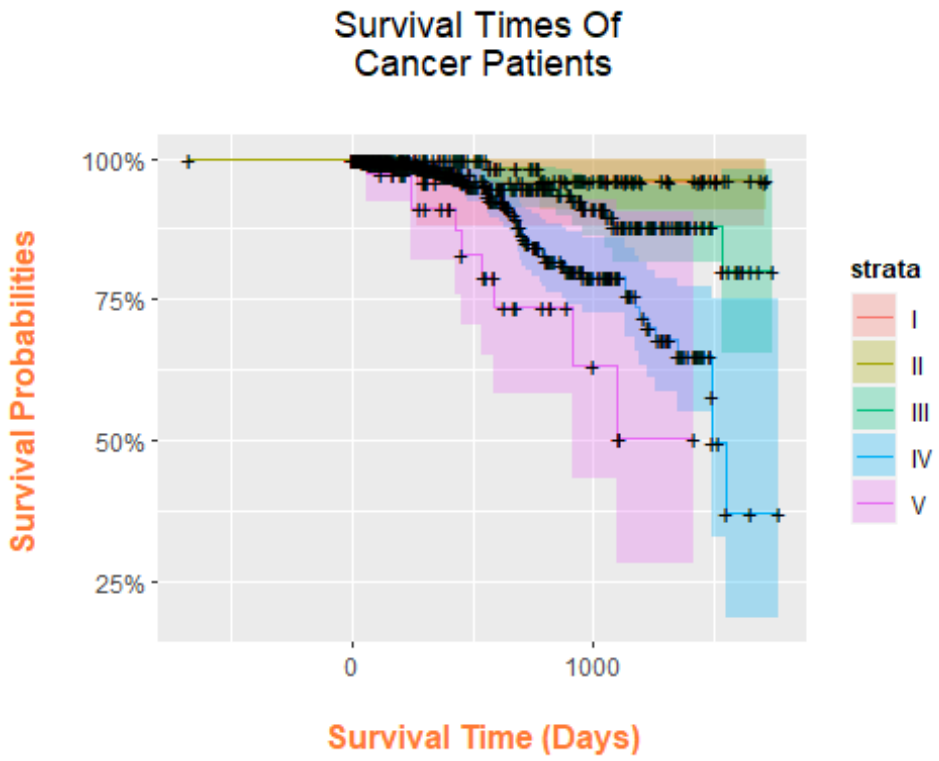
HR of t_stageV vs. t_stageI is 1.9858

ulcer 3.9678

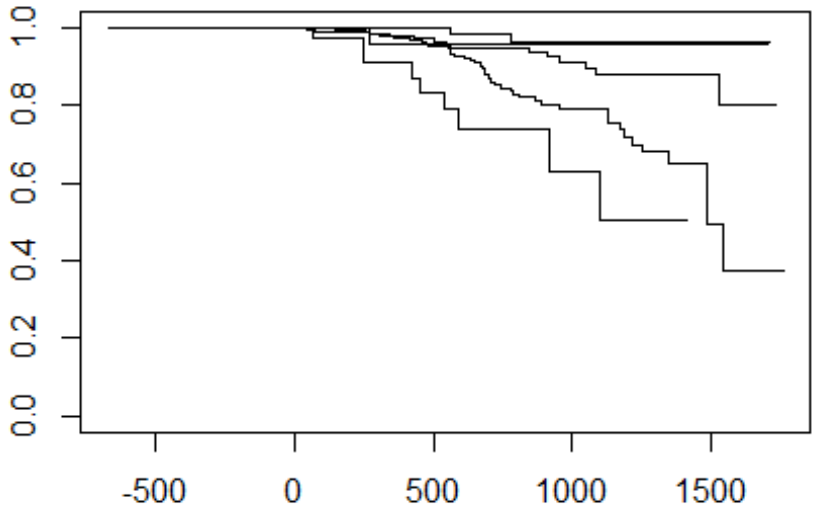
HR of ulcer=1 vs. ulcer=0 is 3.9678

b) Assess the proportional hazard assumption for each of the independent variables. What conclusion do you make in terms of the applicability of the Cox-PH model.

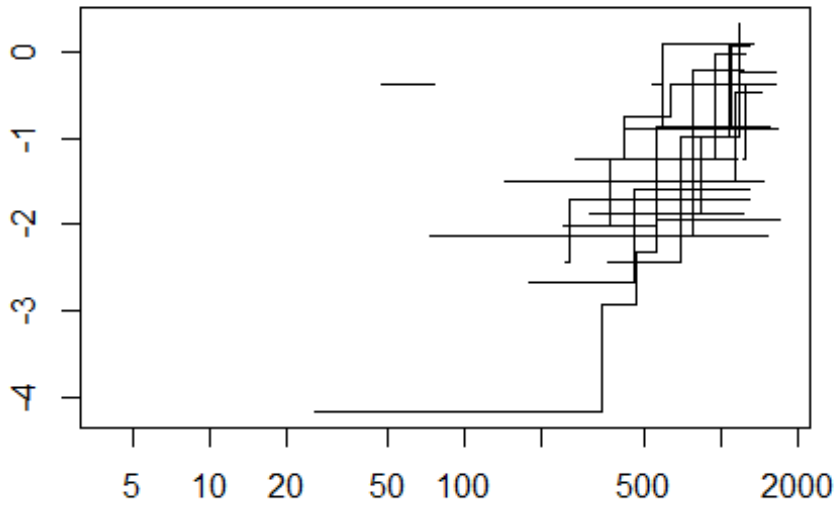
```
m2 <- survfit(Surv(f_time,status)~t_stage, data=pd)


autoplot(m2) +
 labs(x = "\n Survival Time (Days) ", y = "Survival Probabilities \n",
 title = "Survival Times Of \n Cancer Patients \n") +
 theme(plot.title = element_text(hjust = 0.5),
 axis.title.x = element_text(face="bold", colour="#FF7A33", size = 12),
 axis.title.y = element_text(face="bold", colour="#FF7A33", size = 12),
 legend.title = element_text(face="bold", size = 10))
```




```
plot(m2)
```




```
m3 <- survfit(Surv(f_time,status)~thick, data=pd)
plot(m3,fun = "cloglog")
```



ulcer

The graphical test of ulcer can be done by graph plotted in Q1, there is no intersection of the curves, however, the difference between curves increase as time increases, so we conclude ulcer is not PH.

t_stage

The graphical test of t_stage indicate a violation of PH assumption by mutiple intersections of curves i.e. non-parallelism.

thick

The graphical test of thick is on-applicable since thick is a continuous variable.

Since graphical test is subjective and not convenient for variables that have multiple categories or is continuous, we apply GOF.

Apply goodness of fit test

```
cox1 <- coxph(Surv(f_time,status)~t_stage, data=pd)
cox2 <- coxph(Surv(f_time,status)~ulcer, data=pd)
cox3 <- coxph(Surv(f_time,status)~thick, data=pd)
cox.zph(cox1,transform = "identity")

##         chisq df    p
## t_stage  3.44  4 0.49
## GLOBAL   3.44  4 0.49

cox.zph(cox2,transform = "identity")

##        chisq df    p
## ulcer  0.995  1 0.32
## GLOBAL 0.995  1 0.32
```

```r
cox.zph(cox3,transform = "identity")
```

```
##        chisq df   p
## thick   1.63  1 0.2
## GLOBAL  1.63  1 0.2
```

```r
cox.zph(cox1,transform = "rank")
```

```
##         chisq df    p
## t_stage  3.54  4 0.47
## GLOBAL   3.54  4 0.47
```

```r
cox.zph(cox2,transform = "rank")
```

```
##        chisq df   p
## ulcer  0.702  1 0.4
## GLOBAL 0.702  1 0.4
```

```r
cox.zph(cox3,transform = "rank")
```

```
##        chisq df    p
## thick   1.38  1 0.24
## GLOBAL  1.38  1 0.24
```

```r
cox.zph(cox1,transform = "km")
```

```
##         chisq df    p
## t_stage  3.79  4 0.43
## GLOBAL   3.79  4 0.43
```

```r
cox.zph(cox2,transform = "km")
```

```
##        chisq df    p
## ulcer  0.853  1 0.36
## GLOBAL 0.853  1 0.36
```

```r
cox.zph(cox3,transform = "km")
```

```
##        chisq df   p
## thick   1.65  1 0.2
## GLOBAL  1.65  1 0.2
```

```r
cox4 <- coxph(Surv(f_time,status)~t_stage+ulcer+thick, data=pd)
cox.zph(cox4,transform = "identity")
```

```
##         chisq df    p
## t_stage 3.187  4 0.53
## ulcer   0.945  1 0.33
## thick   0.533  1 0.47
## GLOBAL  6.534  6 0.37
```

```r
cox.zph(cox4,transform = "rank")
```

```
##         chisq df    p
## t_stage 3.904  4 0.42
## ulcer   0.624  1 0.43
## thick   0.288  1 0.59
## GLOBAL  6.277  6 0.39
```

```
cox.zph(cox4,transform = "km")
```

```
##         chisq df    p
## t_stage 2.959  4 0.56
## ulcer   0.836  1 0.36
## thick   0.632  1 0.43
## GLOBAL  6.175  6 0.40
```

According to the output, p-value of each model with different transform is large, which indicates that all 3 variables can be assumed PH and Cox PH Model can be applied to this data set.

```
cox4 <- coxph(Surv(f_time,status)~t_stage+ulcer+thick, data=pd)
cox.zph(cox4)
```

```
##         chisq df    p
## t_stage 2.959  4 0.56
## ulcer   0.836  1 0.36
## thick   0.632  1 0.43
## GLOBAL  6.175  6 0.40
```

After checking for PH assumption of model containing all 3 variables, large p-values indicate that Cox-PH Model is applicable for this data set.

Q3

a). Fit a parametric survival model using the given data (use all three independent variables). Find the distribution that best fits the data.

Fitting parametric data requires non-empty data set and follow-up time should be positive.

```
#consider only positive follow-up time
newpd <- pd %>% filter(f_time > 0)
#remove all empty t_stage and ulcer
pd1 <- newpd %>% filter(!is.na(t_stage))%>% filter(!is.na(ulcer))
```

Distribution Selection

```
library(MASS)

library(flexsurv)
distList <- c("weibull", "exp", "gamma", "lnorm","llogis","gompertz")

# Create a function fits each distribution and extracts AIC, BIC
fit_dist <- function(dist) {
  tmp <- flexsurvreg(Surv(f_time,status)~t_stage+thick+ulcer, data=pd1, dist = dist)
  c(AIC(tmp), BIC(tmp), as.numeric(logLik(tmp)))
}

# Apply above function to each distribution in the distList
results_list <- lapply(distList, fit_dist)

# Convert the above list of results to a data frame
results_df <- data.frame(t(matrix(unlist(results_list), nrow = 3)))
colnames(results_df) <- c("aic", "bic")
rownames(results_df) <- distList
```

```r
# Find the distribution with the lowest AIC, BIC, and logLik values
bestFitAIC <- rownames(results_df)[which.min(results_df$aic)]
bestFitBIC <- rownames(results_df)[which.min(results_df$bic)]

bestFitAIC
```

```
## [1] "llogis"
```

```r
bestFitBIC
```

```
## [1] "llogis"
```

Based on AIC and BIC values, conclude that log-logistic Model fits the data set best.

b) You can remove independent variables from the model if you find necessary. Find the best model and interpret the model outputs.

```r
new_m <- survreg(Surv(f_time,status)~t_stage+ulcer+thick, data=pd1,dist = "loglogistic")
stepAIC(new_m)
```

```
## Start:  AIC=908.19
## Surv(f_time, status) ~ t_stage + ulcer + thick
##
##           Df    AIC
## - thick    1 906.20
## <none>       908.19
## - t_stage  4 909.42
## - ulcer    1 924.37
##
## Step:  AIC=906.2
## Surv(f_time, status) ~ t_stage + ulcer
##
##           Df    AIC
## <none>       906.20
## - t_stage  4 911.63
## - ulcer    1 925.86

## Call:
## survreg(formula = Surv(f_time, status) ~ t_stage + ulcer, data = pd1,
##     dist = "loglogistic")
##
## Coefficients:
## (Intercept)   t_stageII  t_stageIII   t_stageIV    t_stageV       ulcer
##   8.1779458   1.0087358   0.2749624  -0.1295343  -0.4310478  -0.7952020
##
## Scale= 0.502821
##
## Loglik(model)= -446.1   Loglik(intercept only)= -468.2
##  Chisq= 44.13 on 5 degrees of freedom, p= 2.18e-08
## n= 668
```

Output generated by stepAIC suggest that thick variable can be dropped from the model and the model will only contain t_stage and ulcer

```r
best_fit_mod <- survreg(Surv(f_time,status)~t_stage+ulcer, data=pd1,dist = "loglogistic")
summary(best_fit_mod)
```

```
##
## Call:
## survreg(formula = Surv(f_time, status) ~ t_stage + ulcer, data = pd1,
##     dist = "loglogistic")
##              Value Std. Error     z       p
## (Intercept)  8.178      0.550 14.86 < 2e-16
## t_stageII    1.009      0.753  1.34    0.18
## t_stageIII   0.275      0.573  0.48    0.63
## t_stageIV   -0.130      0.556 -0.23    0.82
## t_stageV    -0.431      0.605 -0.71    0.48
## ulcer       -0.795      0.187 -4.25 2.1e-05
## Log(scale)  -0.688      0.112 -6.14 8.4e-10
##
## Scale= 0.503
##
## Log logistic distribution
## Loglik(model)= -446.1   Loglik(intercept only)= -468.2
##  Chisq= 44.13 on 5 degrees of freedom, p= 2.2e-08
## Number of Newton-Raphson Iterations: 16
## n= 668
```

t_stageII 1.009

Having t_stageII accelerates the time to event by a factor of exp(1.009) = 2.742857 compared to survival time of t_stage I

t_stageIII 0.275

Having t_stageIII accelerates the time to event by a factor of exp(0.275) = 1.316531 compared to survival time of t_stage I

t_stageIV -0.130

Having t_stageIV accelerates the time to event by a factor of exp(-0.13) = 0.8780954 compared to survival time of t_stage I

t_stageV -0.431

Having t_stageV accelerates the time to event by a factor of exp(-0.431) = 0.6498589 compared to survival time of t_stage I

ulcer -0.795

Having ulcer=1 accelerates the time to event by a factor of exp(-0.795) = 0.4515812 compared to survival time of ulcer=0

Q4 Summary

```
max(pd$f_time)
```

```
## [1] 1761
```

The maximum follow-up time is 1761 and KM curve shows that the probability of surviving longer than 1761 days is 0.65. Since there is no upper bound of survival time(i.e. the upper bound of survival time is infinity), and the probability of surviving longer than 1761 days is 0.65 > 0.5, the median survival time is infinity, therefore, we cannot calculate an exact 95%-CI for median survival time.

KM curves generated w.r.t. ulcer=0 and ulcer=1 shows that ulcer=1 generally have shorter survival time, and this complies to the result generated in Cox-PH model and the parametric model that follows log-logistic distribution, the positive coefficients of coxph model suggest that ulcer = 1 has higher hazard and the negative coefficient of parametric model with log-logistic distribution indicates that ulcer=1 contracts survival time.

Cox-PH Model that fits the data best should contain t_stage and ulcer based on AIC, the result complies to the KM curves, ulcer = 1 increases the hazard and t_stage II has lowest hazard.

While checking for PH assumption of each variables, the KM curves and GOF method generate different conclusions. Ulcer and t_stage are categorical variables, their KM curves are non-parallel and their are intersections for t_stage curves, subjectively we conclude that they do not follow PH assumption. GOF result suggest that when considering ulcer,t_stage and thick respectively in models with single variable, they all follow PH assumption. In the model containing all 3 variables together, it also follows PH assumption. Therefore, Cox-PH model can be applied to this data set.

The parametric model with log-logistic distribution fits the data set best based on comparison of AIC and BIC, and the model contains only the ulcer and t_stage variables. Holding ulcer the fixed, t_stage II has longest survival time, holding t_stage fixed, ulcer = 0 has longer survival time. However, the large p-values of t_stage suggest that t_stage is not very influential.

The parametric model and Cox-PH Model summary results have different signs for coefficients because survreg() generates AFT model whereas Cox-PH model focuses on hazards, however, both these 2 models concludes same results of how t_stage and ulcer affect survival time. Cox-PH Model have lower AIC, Cox-PH model does not force the data to follow a certain distribution, I suppose that Cox-PH model is better applied.