Ch5Report

Madi Kassymbekov

Question 1

There are two records per person, one for cancer recurrence and one for death. The variables are the following: id: subject id study: 1 for all patients rx: Treatment: Obs (no chemio), Lev (Levamisole), Lev+5-FU (Levamisol 5FU) sex: 1=male; 0=female age: age in years of the patient at baseline obstruct: obstruction of colon by tumour (1=yes / 0=no) perfor: perforation of colon (1=yes / 0=no) adhere: adherence to nearby organs (1=yes / 0=no) nodes: number of lymph nodes with detectable cancer time: days until event or censoring status: censoring status differ: differentiation of tumour (1=well, 2=moderate, 3=poor) extent: Extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous) surg: time from surgery to registration (0=short, 1=long) node4: more than 4 positive lymph nodes etype: event type: 1=cancer recurrence, 2=death

1) Prepare the dataset in order to consider the following:

- We'll consider only the time-to-death as the event of interest (time to recurrence will not be used)
- Variables differ and extent will be considered as numeric since they are ordered
- The study variable will not be used
- The first 500 subjects will be used for training and the remaining subjects will be used as a test set.

```
library(survival)
#Only time-to-death rows
colondata <- colon[colon$etype==2,]
colondata$study <- NULL
colondata$etype <- NULL
trainset <- colondata[1:500,]
testset <- colondata[501:929,]
```

2) Compute basic descriptive statistics on the training dataset:

i) What is the censoring rate ?

```
#censoring rate
censorRate <- nrow(trainset[trainset$status==1,])/nrow(trainset)
print(sprintf("Censoring Rate: %f", censorRate))
```

```
## [1] "Censoring Rate: 0.508000"
```

Answer: Censoring Rate of training dataset is 0.508.

ii) What is the mean/median/min/max values of the survival time ?

```
#mean survival time
mean <- mean(trainset$time)
#median survival time
median <- median(trainset$time)
#min survival time
min <- min(trainset$time)
#max survival time
max <- max(trainset$time)
```

```
results <- data.frame("Overall Descriptive Statistics (Survival Time in days)",
mean, median, min, max)
names(results) <- c("Name", "Mean", "Median", "Min", "Max")
knitr::kable(results, caption="mean/median/min/max values of the survival time in days")
```

Table 1: mean/median/min/max values of the survival time in days

Name	Mean	Median	Min	Max
Overall Descriptive Statistics (Survival Time in days)	1784.834	2189	23	3329

Answer: For survival time in training dataset mean is 1784.834 days, median is 2189 days, minimum is 23 days and maximum is 3329 days.

iii) Re-compute the mean/median/min/max values of the survival time, for each value of the status variable. Explain and interpret the difference in the distributions.

```
#survival time statistics for status = 1 (censored)
meanCensor <- mean(trainset[trainset$status==1, "time"])
medianCensor <- median(trainset[trainset$status==1, "time"])
minCensor <- min(trainset[trainset$status==1, "time"])
maxCensor <- max(trainset[trainset$status==1, "time"])

#survival time statistics for status = 0 (non-censored)
meanNoCensor <- mean(trainset[trainset$status==0, "time"])
medianNoCensor <- median(trainset[trainset$status==0, "time"])
minNoCensor <- min(trainset[trainset$status==0, "time"])
maxNoCensor <- max(trainset[trainset$status==0, "time"])

results[nrow(results) + 1,] = list(Name="Survival time stats for censored status=1",
                                   meanCensor, medianCensor,minCensor,maxCensor)
results[nrow(results) + 1,] = list(Name="Survival time stats for non-censored status=0",
                                   meanNoCensor, medianNoCensor,minNoCensor,maxNoCensor)
knitr::kable(results[2:3,], caption="Survival Time Statistics for each status")
```

Table 2: Survival Time Statistics for each status

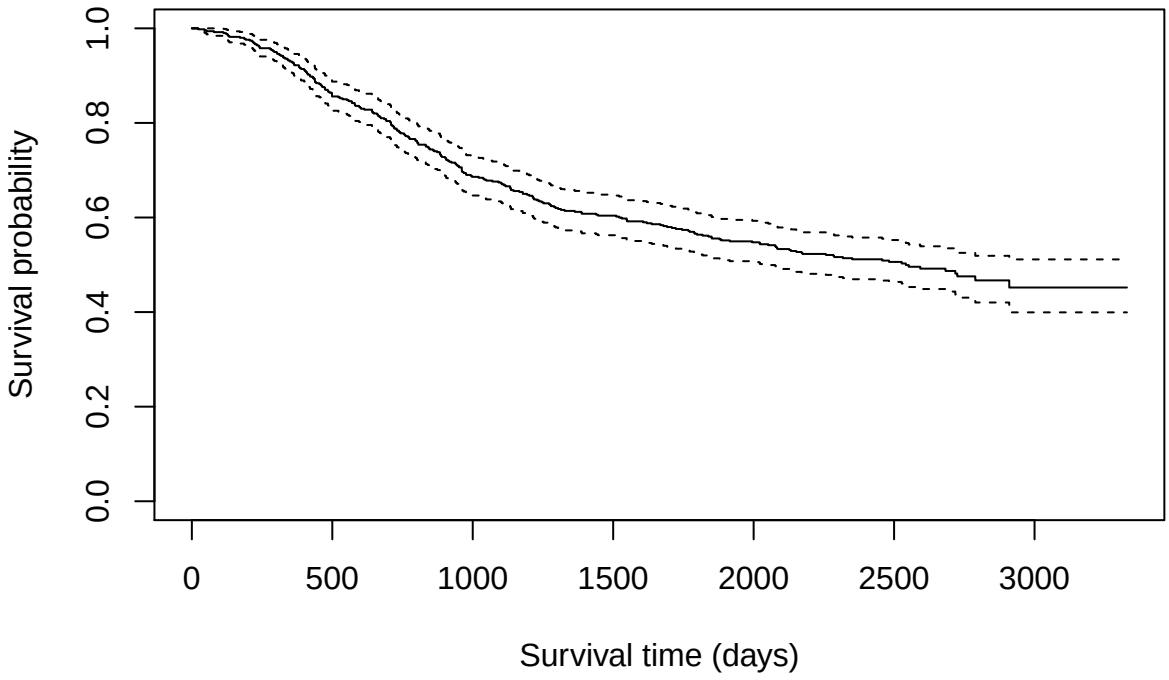
	Name	Mean	Median	Min	Max
2	Survival time stats for censored status=1	996.6142	845.5	23	2910
3	Survival time stats for non-censored status=0	2598.6870	2584.0	1279	3329

Answer: Survival time statistics for each status are present above. It should be noted that each stat value is much higher for non-censored with status = 0 than for censored.

3) Let's see what the global estimated survival curve looks like...

i) Draw a Kaplan-Meyer (KM) estimate of the survival curve for the training data.

```
kmfit=survfit(Surv(time, status) ~ 1, type="kaplan-meier", conf.type="log", data=trainset)
plot(kmfit, xlab="Survival time (days)", ylab="Survival probability")
```



ii) What is the estimated probability to survive at least 400 days ?

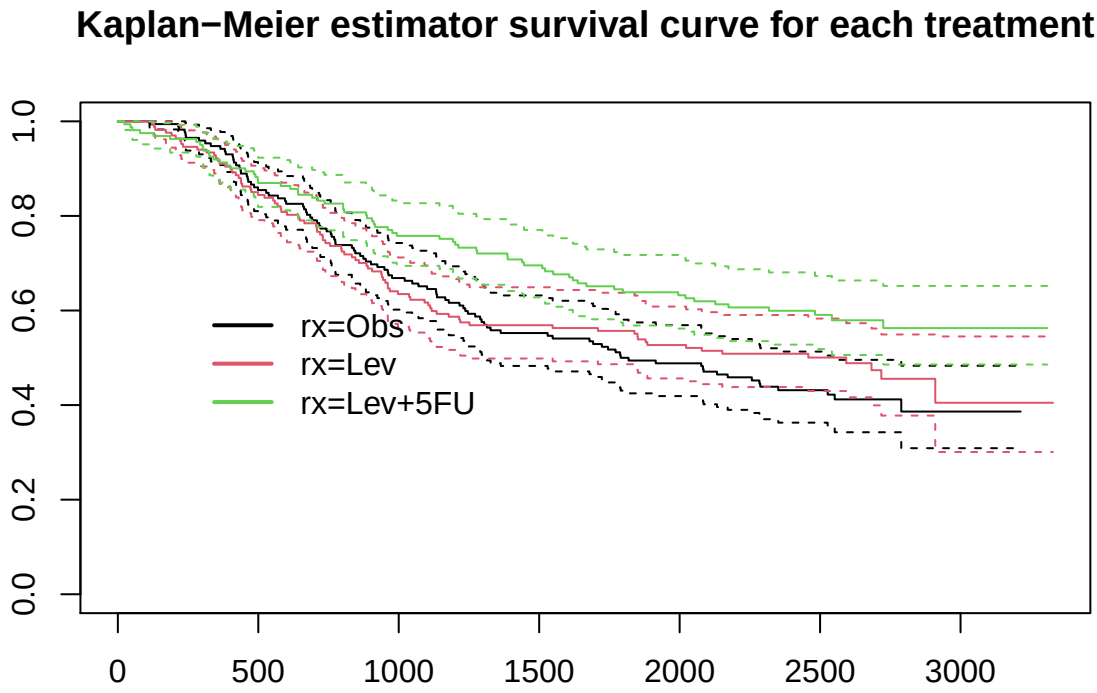
```
summary(kmfit, times = 400)
```

```
## Call: survfit(formula = Surv(time, status) ~ 1, data = trainset, type = "kaplan-meier",
##     conf.type = "log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   400    457      44    0.912  0.0127        0.888        0.937
```

Probability to survive 400 days is 0.912 based on kaplan-meier estimator summary.

iii) Draw the KM estimate of the survival curve for each type of treatment. Comment.

```
treatkmfit=survfit(Surv(time, status) ~ rx, type="kaplan-meier", conf.type="log",
data=trainset)
plot(treatkmfit, col=1:3, conf.int=TRUE)
legend(240, .65, c("rx=Obs", "rx=Lev", "rx=Lev+5FU"),
lty=c(1,1,1), col=c(1,2,3), bty='n', lwd=2)
title("Kaplan-Meier estimator survival curve for each treatment")
```



Answer: Looking at KM curves, it is obvious that treatments Lev+5FU increases the survival probability across longer periods. Treatment Lev is not stable across longer periods and performs similarly or worse compared to the subjects with no treatment.

iv) Explain why the confidence interval for the survival curve is larger towards the end of the curve Answer: As the survival period increases, number of observations decreases which affects/worsens the estimation precision of the confidence interval.

4) Perform a log-rank test to compare the survival curves between treatment. Comment.

```
survdiff(formula=Surv(time, status)~rx, data=trainset)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ rx, data = trainset)
##
##               N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=Obs      172      100     84.1     3.018     4.520
## rx=Lev      167       87     81.6     0.358     0.529
## rx=Lev+5FU 161       67     88.3     5.154     7.925
##
##  Chisq= 8.6  on 2 degrees of freedom, p= 0.01
```

Answer: Chisquare statistic based on log-rank test is 8.6 and p-value is 0.01 by which we reject the null hypothesis that there is no difference in survival times between treatments.

5) Use a Cox proportional hazard model to evaluate the effect of the treatment (use only this variable in the model).

i) Interpret the two coefficients of the model (on the exponential scale)
ii) Make a link between some parts of the R output and the previous question

```
fitcox=coxph(Surv(time, status)~rx, data=trainset)
summary(fitcox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = trainset)
##
##   n= 500, number of events= 254
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.1095    0.8962   0.1467 -0.747  0.45511
## rxLev+5FU -0.4511    0.6369   0.1581 -2.854  0.00432 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.8962      1.116    0.6723    1.1947
## rxLev+5FU    0.6369      1.570    0.4672    0.8682
##
## Concordance= 0.543  (se = 0.017 )
## Likelihood ratio test= 8.85  on 2 df,   p=0.01
## Wald test            = 8.44  on 2 df,   p=0.01
## Score (logrank) test = 8.55  on 2 df,   p=0.01
```

i) Based on the Cox proportional hazard model summary, if subject is given a dose of rxLev treatment and everything else remains fixed, the risk of time to death is multiplied by 0.8962 and therefore risk of death decreases compared with no treatment. Similarly for rxLev+5FU, risk of death is multiplied by 0.6369 and therefore risk of death decreases by even more compared to rxLev everything else remaining fixed.

ii) Cox proportional hazard model summary also provides logrank test results automatically, which values are the same as performing it with survdiff. P-values are the same, but chisq is rounded to one decimal point in survdiff while in coxph it is up to two decimal points. Cox model also proved the log-rank test that there is a difference in survival times between treatments and Lev+5FU provides lowest risk of death compared to others.

6) Use a Cox proportional hazard to evaluate the effect of all variables.

i) Predict the risk of death for all subjects in the test set. Interpret this risk for the 1st subject (id=501)

```
fitcox=coxph(Surv(time, status)~ . -id, data=trainset)
summary(fitcox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ . - id, data = trainset)
##
##   n= 477, number of events= 241
##    (23 observations deleted due to missingness)
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## rxLev     -0.091319  0.912726  0.153257 -0.596 0.551271
## rxLev+5FU -0.483731  0.616479  0.163351 -2.961 0.003064 **
## sex        0.071514  1.074134  0.130579  0.548 0.583917
## age        0.018110  1.018275  0.006093  2.972 0.002956 **
## obstruct   0.150923  1.162907  0.162975  0.926 0.354422
## perfor    -0.307705  0.735132  0.425597 -0.723 0.469682
## adhere     0.113156  1.119806  0.177378  0.638 0.523516
## nodes      0.024375  1.024675  0.019281  1.264 0.206146
```

```
## differ     0.284214  1.328717  0.137363  2.069 0.038539 *
## extent     0.540453  1.716785  0.156237  3.459 0.000542 ***
## surg       0.075297  1.078204  0.144019  0.523 0.601095
## node4      0.841784  2.320504  0.188558  4.464 8.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9127     1.0956    0.6759    1.2325
## rxLev+5FU    0.6165     1.6221    0.4476    0.8491
## sex          1.0741     0.9310    0.8316    1.3874
## age          1.0183     0.9821    1.0062    1.0305
## obstruct     1.1629     0.8599    0.8449    1.6006
## perfor       0.7351     1.3603    0.3192    1.6929
## adhere       1.1198     0.8930    0.7910    1.5854
## nodes        1.0247     0.9759    0.9867    1.0641
## differ       1.3287     0.7526    1.0151    1.7392
## extent       1.7168     0.5825    1.2639    2.3319
## surg         1.0782     0.9275    0.8130    1.4298
## node4        2.3205     0.4309    1.6035    3.3580
##
## Concordance= 0.679  (se = 0.017 )
## Likelihood ratio test= 93.56  on 12 df,   p=1e-14
## Wald test            = 95.9  on 12 df,   p=4e-15
## Score (logrank) test = 101.2  on 12 df,   p=3e-16
```

```
predcoxrisk=predict(fitcox, newdata=testset, type="risk")
id_501_risk <-predcoxrisk[1]
```

Answer: Compared to average person, risk of death for subject 501 is 0.7239861.

ii) Explain why some missing values are generated for the predicted risks

```
testset$risk <- predcoxrisk

testset[is.na(testset$risk)==TRUE,]
```

```
##       id      rx sex age obstruct perfor adhere nodes status differ extent surg
## 1003 502 Lev+5FU   0  71        0      0      0    NA      1      2      3    0
## 1011 506     Lev   1  57        0      0      0     3      0     NA      3    0
## 1043 522     Obs   0  72        0      0      0    NA      0      2      3    0
## 1145 573     Obs   1  56        0      0      0     1      1     NA      4    0
## 1179 590     Lev   1  54        0      0      1    NA      1      2      3    1
## 1217 609 Lev+5FU   1  66        1      0      0    NA      1      3      1    1
## 1255 628 Lev+5FU   1  59        0      0      0     2      0     NA      3    0
## 1271 636 Lev+5FU   0  71        0      0      0    NA      0      1      3    0
## 1289 645     Lev   1  66        0      0      1     1      1     NA      4    0
## 1343 672     Obs   1  73        0      0      1     3      0     NA      2    0
## 1385 693     Obs   1  38        1      0      0     2      1     NA      4    0
## 1471 736     Obs   1  57        0      0      0    NA      0      2      3    0
## 1475 738     Obs   1  60        0      0      0     2      0     NA      3    0
## 1541 771     Lev   0  65        1      0      0    NA      1      3      3    1
## 1573 787     Obs   0  57        0      0      0    NA      1      2      3    0
## 1629 815     Obs   0  56        0      0      0     2      0     NA      1    0
## 1637 819 Lev+5FU   1  31        1      0      0    NA      1      1      3    1
## 1713 857 Lev+5FU   0  54        0      0      0     3      0     NA      2    0
```

```
##      node4 time risk
## 1003     1 1273   NA
## 1011     0 2264   NA
## 1043     1 2257   NA
## 1145     0 1884   NA
## 1179     1 1061   NA
## 1217     1  583   NA
## 1255     0 2120   NA
## 1271     1 2290   NA
## 1289     0 1540   NA
## 1343     0 2210   NA
## 1385     0  469   NA
## 1471     1 2191   NA
## 1475     0 1856   NA
## 1541     1  444   NA
## 1573     1  201   NA
## 1629     0 1990   NA
## 1637     1  643   NA
## 1713     0 2070   NA
```

Answer: Persons with missing risk had missing covariates and therefore risk could not be estimated.

iii) Give the list of the four subjects with the highest risk. By looking at their characteristics, make a link with the results of the Cox model you just fitted.

```
toprisk <- testset[order(testset$risk, decreasing = TRUE, na.last = TRUE),]
toprisk[1:4,]
```

```
##       id      rx sex age obstruct perfor adhere nodes status differ extent surg
## 1643 822     Lev   1  60        1      0      0    14      1      2      4    1
## 1045 523 Lev+5FU   0  77        0      0      0    10      0      3      4    0
## 1561 781     Obs   1  66        0      0      0     7      1      2      4    0
## 1813 907     Obs   0  68        0      0      0    12      1      3      3    1
##      node4 time     risk
## 1643     1  512 6.593639
## 1045     1 2381 5.422448
## 1561     1  259 5.415356
## 1813     1  887 4.927739
```

Answer: Based on the cox model summary,three most impactful covariates for risk of death increase are node4 (more than 4 lymph nodes), extent (local spread extent) and differ (differentiation of tumour) with coefficients of 2.3205, 1.7168 and 1.3287 respectively. Top4 risk of death ids all have more than 4 lymph nodes, high possible extent of 3 and 4 (serosa and contigious structures) and differ value of 2 and 3 (moderate to poor tumour) which all contributed to the high risk values even thought top 2 ids had Lev and Lev+5FU treatments.

7) Now, fit an accelerated Failure Time (AFT) model using all variables, assuming a lognormal distribution.

i) Interpret the coefficient of the age variable (on the exponential scale)

```
fitaft = survreg(Surv(time, status)~. -id, data=trainset, dist="lognormal")
summary(fitaft)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ . - id, data = trainset,
##     dist = "lognormal")
##                Value Std. Error     z       p
```

```
## (Intercept) 11.92575    0.69373 17.19 < 2e-16
## rxLev        0.03146    0.15625  0.20 0.84043
## rxLev+5FU    0.29099    0.16092  1.81 0.07056
## sex         -0.00930    0.13090 -0.07 0.94333
## age         -0.01741    0.00587 -2.97 0.00300
## obstruct    -0.22500    0.16376 -1.37 0.16946
## perfor       0.50062    0.41173  1.22 0.22403
## adhere      -0.07211    0.18104 -0.40 0.69040
## nodes       -0.02461    0.02341 -1.05 0.29318
## differ      -0.35923    0.13355 -2.69 0.00715
## extent      -0.71550    0.16439 -4.35 1.3e-05
## surg        -0.12077    0.14471 -0.83 0.40395
## node4       -0.80202    0.20739 -3.87 0.00011
## Log(scale)   0.22957    0.05027  4.57 5.0e-06
##
## Scale= 1.26
##
## Log Normal distribution
## Loglik(model)= -2149.6   Loglik(intercept only)= -2197.7
##  Chisq= 96.2 on 12 degrees of freedom, p= 3.1e-15
## Number of Newton-Raphson Iterations: 4
## n=477 (23 observations deleted due to missingness)
```

Answer: When the age increases by one, then the average time of death is multiplied by exp(-0.017)=0.983, so decreases by a little bit.

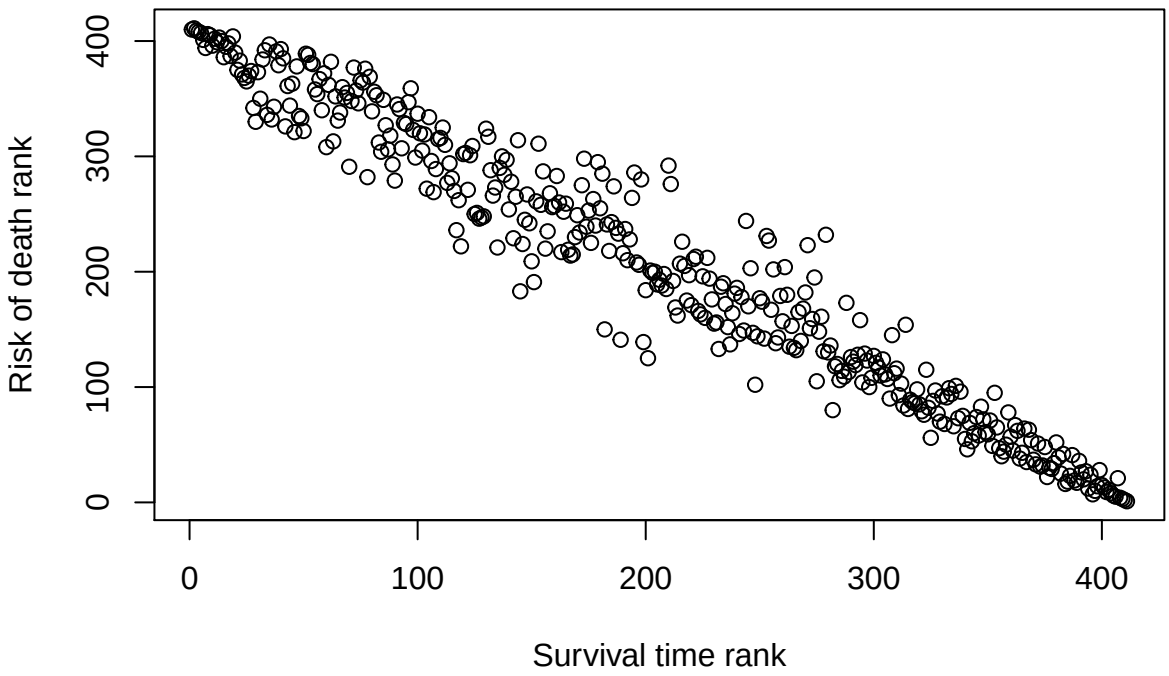
ii) On the test data, compute the predicted median survival time. Compare it with the actual survival time and comment.

```
predaft = predict(fitaft, newdata=testset, type="response")
testset$survtime <- predaft
testsetNoNa <- na.omit(testset)
medianActualTime <- median(testsetNoNa$time)
medianAFTTime <- median(testsetNoNa$survtime)
aftMAE <- mean(abs(testsetNoNa$survtime-testsetNoNa$time))
aftMSE <- mean((testsetNoNa$survtime-testsetNoNa$time)^2)
```

Answer: As 18 rows of test set had missing covariates, these rows were excluded from performance comparison. Median survival time predicted by AFT is 2714.2843694, while actual median survival time is 1896. AFT model overestimated the survival time by almost 1000 days which is too promising to be true compared to actual survival times which is proved by huge values of MSE is 7.9050427×10^6 and MAE is 1755.4688491.

iii) Plot the ranks of the subjects according to their predicted risk from the Cox model versus the rank computed from the predicted survival time. Comment.

```
topsurvival <- testsetNoNa[order(testsetNoNa$survtime, decreasing = TRUE),]
topsurvival$rank <- seq_along(topsurvival[,1])
toprisk <- testset[order(testset$risk, decreasing = TRUE),]
toprisk <- na.omit(toprisk)
toprisk$rank <- seq_along(toprisk[,1])
topsurvival <- topsurvival[order(topsurvival$id, decreasing = FALSE),]
toprisk <- toprisk[order(toprisk$id, decreasing = FALSE),]
testsetNoNa$survrank <- topsurvival$rank
testsetNoNa$riskrank <- toprisk$rank
plot(testsetNoNa$survrank, testsetNoNa$riskrank, xlab="Survival time rank",
ylab="Risk of death rank")
```



Answer: Based on the plot of survival time rank vs risk of death rank we can observe that there is more or less an agreement between cox and aft model predictions where a linear relationship can be seen. Survival time increases when the risk of death decreases.

Question 2

Which of the following statements about survival analysis concepts is TRUE?

A. One of the disadvantages of Cox model that there is a need to select probability distribution for a target variable.

Answer: False. Compared to other models, Cox does not require target distribution however it can complicate computation of predictions.

B. Kaplan-Meier estimator is a parametric statistic to estimate the survival function from given time data.

Answer: False. Kaplan-Meier estimator is in fact a non-parametric statistic. It does not assume any underlying probability distribution rather conditional probabilities are used at each time t.

C. When analyzing survival data one of the difficulties is that not all events can be observed.

Answer: True. Some subjects drop out before the end of experiment, some subjects survive well beyond the end of experiment time, etc. Therefore, censoring is used to account for these subjects through the observed time t to have some general information on such subjects.

D. AFT model can provide predictions for median survival regardless of the distribution of a target variable.

Answer: False. AFT can provide estimated median survival time only to gaussian, lognormal, loglogistic and logistic distributions.