Description of the data:

The samples are taken from the colorectal cancer patients. All the patients have undergone sugery as a primary treatment. Apart from the primary treatment some patients have taken chemotherapy and radiotherapy or both. Here the time to event is “how many months the patients have survived without the disease after the treatment”.

```
dat <- load("CRC_226_GSE14333.RData")

dim(clinical_data)

## [1] 226   9

head(clinical_data)

##    sampleID location dukes_stage age_diag gender dfs_time dfs_event adjXRT
## 1 GSM358341    Right           A       78      M     3.64         1      N
## 2 GSM358342   Rectum           A       53      F    14.53         0      N
## 3 GSM358343     Left           A       80      F    16.47         1      N
## 4 GSM358344     Left           A       58      M    19.75         1      N
## 5 GSM358345     Left           A       81      M    20.02         1      N
## 6 GSM358346    Right           A       57      M    23.96         1      N
##   adjCTX
## 1      N
## 2      N
## 3      N
## 4      N
## 5      N
## 6      N

str(clinical_data)

## 'data.frame':    226 obs. of  9 variables:
##  $ sampleID   : chr  "GSM358341" "GSM358342" "GSM358343" "GSM358344" ...
##  $ location   : Factor w/ 4 levels "Rectum","Colon",..: 4 1 3 3 3 4 3 3 4
4 ...
##  $ dukes_stage: Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ age_diag   : num  78 53 80 58 81 57 63 51 86 76 ...
##  $ gender     : Factor w/ 2 levels "F","M": 2 1 1 2 2 2 1 2 1 2 ...
##  $ dfs_time   : num  3.64 14.53 16.47 19.75 20.02 ...
##  $ dfs_event  : num  1 0 1 1 1 1 0 1 1 1 ...
##  $ adjXRT     : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ adjCTX     : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 2 1 1 ...
```

variable description:

sampleID: Unique id for each individual.

location: Location of the cancer. It is a categorical variable with 4 values namely Colon, Rectum, Right, Left.

Dukes Stage: Classification of cancer. It is a categorical variable with 3 levels A,B,C. "C" being the advanced stage.

age_diag: Age of the patient and it is a continuous variable.

gender: sex of the patient. "F" -> Female "M"->Male.

dfs_time: Disease free survival time in months.

dfs_event: Indicator to indicate whether the event has occurred or censored. 0->censoring, 1->event has occurred.

adjXRT: Says whether the patient has taken radio therapy. has two values "Y"-> Yes and "N"->No.

adjCTX: Says whether the patient has taken radio therapy. has two values "Y"-> Yes and "N"->No.

Summary of the data

Made table for each variable to understand better about the data set and we can see that data set has no missing values and from the histogram of the age we can notice that most of the observations lies between age group 50-80 years.

```
sum(is.na(clinical_data) | clinical_data == "")
```

```
## [1] 0
```

```
table(clinical_data$location, dnn = "Number of observations based on location
of the cancer")
```

```
## Number of observations based on location of the cancer
## Rectum  Colon   Left  Right
##     30      2     93    101
```

```
table(clinical_data$dukes_stage,dnn = "Number of observations based on Dukes
stage of the cancer")
```

```
## Number of observations based on Dukes stage of the cancer
##  A  B  C
## 41 94 91
```

```r
table(clinical_data$gender,dnn = "Number of observations based on Gender")
```

```
## Number of observations based on Gender
##   F   M
## 106 120
```

```r
table(clinical_data$dfs_event, dnn = "Number of observations that have been censored and which experienced the event")
```

```
## Number of observations that have been censored and which experienced the event
##   0   1
##  50 176
```

```r
hist(clinical_data$age_diag, xlab = "Patients age", main = "Histogram of patients age")
```


```r
table(clinical_data$adjXRT=="N" & clinical_data$adjCTX == "N", dnn = "Number of observations underwent just the surgery")
```

```
## Number of observations underwent just the surgery
## FALSE  TRUE
##    88   138
```

```r
table(clinical_data$adjCTX, dnn = "Number of observations underwent Chemo Therapy after surgery")
```

```
## Number of observations underwent Chemo Therapy after surgery
##   N   Y
## 139  87

table(clinical_data$adjXRT, dnn = "Number of observations underwent Radio
Therapy after surgery")

## Number of observations underwent Radio Therapy after surgery
##   N   Y
## 204  22

table(clinical_data$adjXRT=="Y" & clinical_data$adjCTX == "Y", dnn = "Number
of observations underwent Both treatment after surgery")

## Number of observations underwent Both treatment after surgery
## FALSE  TRUE
##   205    21
```

From the below summary we can see median and quartiles for the continuous variable and we can also make sure that all the variable are in same type as described before.

```
summary(clinical_data)

##    sampleID           location   dukes_stage    age_diag     gender
##  Length:226         Rectum: 30   A:41        Min.   :26.00   F:106
##  Class :character   Colon :  2   B:94        1st Qu.:58.00   M:120
##  Mode  :character   Left  : 93   C:91        Median :67.00
##                     Right :101               Mean   :66.03
##                                              3rd Qu.:75.00
##                                              Max.   :92.00
##     dfs_time        dfs_event      adjXRT  adjCTX
##  Min.   :  0.92   Min.   :0.0000   N:204   N:139
##  1st Qu.: 22.28   1st Qu.:1.0000   Y: 22   Y: 87
##  Median : 38.46   Median :1.0000
##  Mean   : 43.52   Mean   :0.7788
##  3rd Qu.: 59.50   3rd Qu.:1.0000
##  Max.   :142.55   Max.   :1.0000
```

Survival Analysis

Question asked: To find the variables that has significance in construction of the model.

converting the disease-free survival time from months to years for the ease of work.

```
library(survival)
clinical_data$test = with(clinical_data, Surv(dfs_time/12,dfs_event))
```

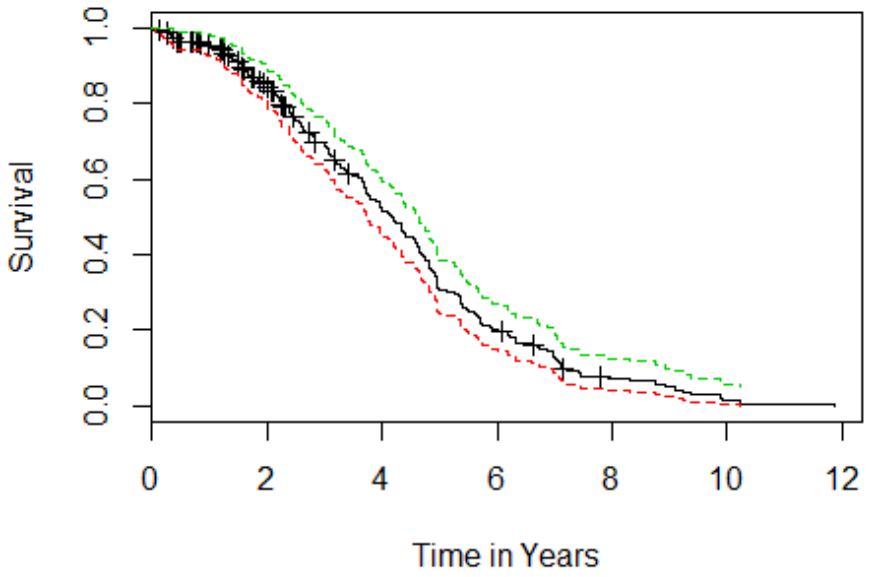
Next step is to check for the trend of the survival curve.

The Kaplan-Meier survival curve shows the cumulative proportion of patients survived over time. The rate of loss of patient is relatively constant over time. The median survival time is 4.15 years. Most of the censored observation are before the median survival time.

```
survfit(test~1, data = clinical_data)

## Call: survfit(formula = test ~ 1, data = clinical_data)
##
##       n  events  median 0.95LCL 0.95UCL
##  226.00  176.00    4.15    3.73    4.64

plot(survfit(test~1, data = clinical_data), col = 1:3, xlab = "Time in
Years", ylab ="Survival", mark.time = TRUE)
```



Now we run the Kaplan-Meier test for all individual variable.

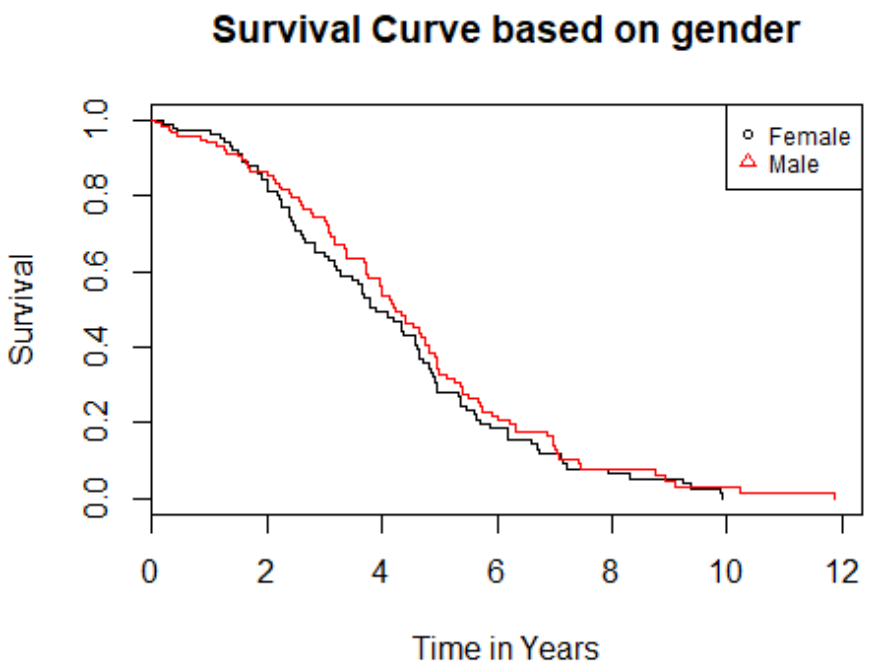
From the graph, we do not see any noticeable difference between the levels of gender and location. We have also confirmed this by running logrank test which gives high p-value. Higher p-value means that we fail to reject null hypothesis which says that there is no significant difference between the levels of the variable.

So, we may omit these variables while building the model.

```r
survfit(test~gender, data = clinical_data)
```

```
## Call: survfit(formula = test ~ gender, data = clinical_data)
##
##             n events median 0.95LCL 0.95UCL
## gender=F 106     84   3.89    3.27    4.66
## gender=M 120     92   4.24    3.75    4.83
```

```r
plot(survfit(test~gender, data = clinical_data), col = 1:2,xlab = "Time in
Years", ylab ="Survival",pch = seq(1,2) )
legend(x = "topright",legend=c("Female","Male"),pch = seq(1,2) ,bty
="o",col=seq(1,2),cex = 0.75)
title("Survival Curve based on gender")
```



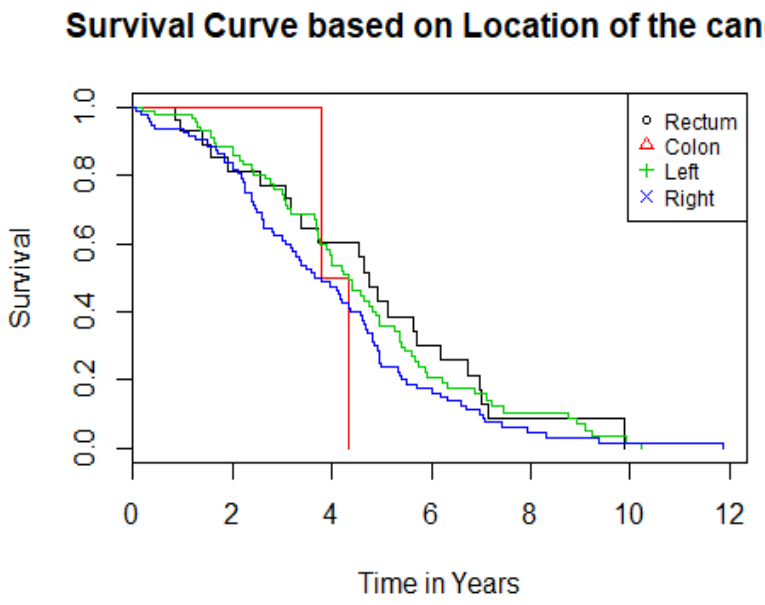

```r
survdiff(test~gender, data = clinical_data)
```

```
## Call:
## survdiff(formula = test ~ gender, data = clinical_data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## gender=F 106       84     78.4     0.406     0.741
## gender=M 120       92     97.6     0.326     0.741
##
##  Chisq= 0.7  on 1 degrees of freedom, p= 0.389
```

```
survfit(test~location, data = clinical_data)
```

```
## Call: survfit(formula = test ~ location, data = clinical_data)
##
##                   n events median 0.95LCL 0.95UCL
## location=Rectum  30     23   4.75    3.37    6.73
## location=Colon    2      2   4.06    3.78      NA
## location=Left    93     68   4.34    3.75    4.96
## location=Right  101     83   3.77    3.17    4.61
```

```
plot(survfit(test~location, data = clinical_data), col = 1:4,xlab = "Time in
Years", ylab ="Survival")
legend(x = "topright",legend=c("Rectum","Colon","Left","Right"),pch =
seq(1,4) ,bty ="o",col=seq(1,4),cex = 0.75)
title("Survival Curve based on Location of the cancer")
```




```
survdiff(test~location, data = clinical_data)
```

```
## Call:
## survdiff(formula = test ~ location, data = clinical_data)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## location=Rectum  30       23    27.29     0.673     0.804
## location=Colon    2        2     1.34     0.321     0.325
## location=Left    93       68    75.42     0.730     1.292
## location=Right  101       83    71.95     1.696     2.913
##
##  Chisq= 3.5  on 3 degrees of freedom, p= 0.323
```

We Perform the Kaplan-Meier test for other variables along with the longrank test to see the significance of the variable in building the model. Since we can't perform logrank test in a continuous variable and age of the patients is a continuous variable. so, we are converting into a categorical variable by dividing the observations into three group (i.e. 0-50,50-80 and 80-inf).

From the below analysis we can say that variable adjXRX (variable indicating whether the patient has taken radio therapy or not) has smaller p-value so we can reject the null hypothesis and accept the alternative one which says that the difference between the two group is significant.

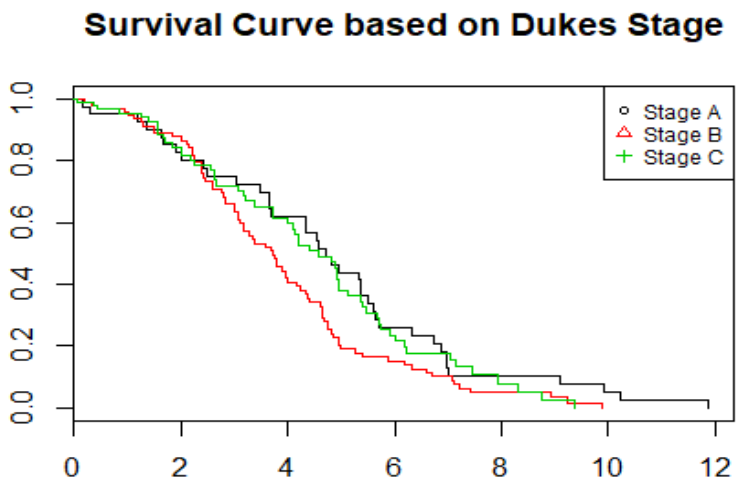
similarly, for the other variable age, dukes stage doesn't show any high significant but in real world situation survival of the cancer patients also depends on the cancer stage so I have decided to consider this variable in the model building.

For adjCTX (variable indicating whether the patient has taken chemo therapy or not) although p-value indicates that the difference between the variable group is not significant.

```
survfit(test~dukes_stage, data = clinical_data)

## Call: survfit(formula = test ~ dukes_stage, data = clinical_data)
##
##                 n events median 0.95LCL 0.95UCL
## dukes_stage=A 41     39   4.71    3.68    5.60
## dukes_stage=B 94     80   3.73    3.17    4.34
## dukes_stage=C 91     57   4.58    3.98    5.36

plot(survfit(test~dukes_stage, data = clinical_data), col = 1:3)
legend(x = "topright",legend=c("Stage A","Stage B","Stage C"),pch = seq(1,3)
,bty ="o",col=seq(1,3),cex = 0.75)
title("Survival Curve based on Dukes Stage")
```



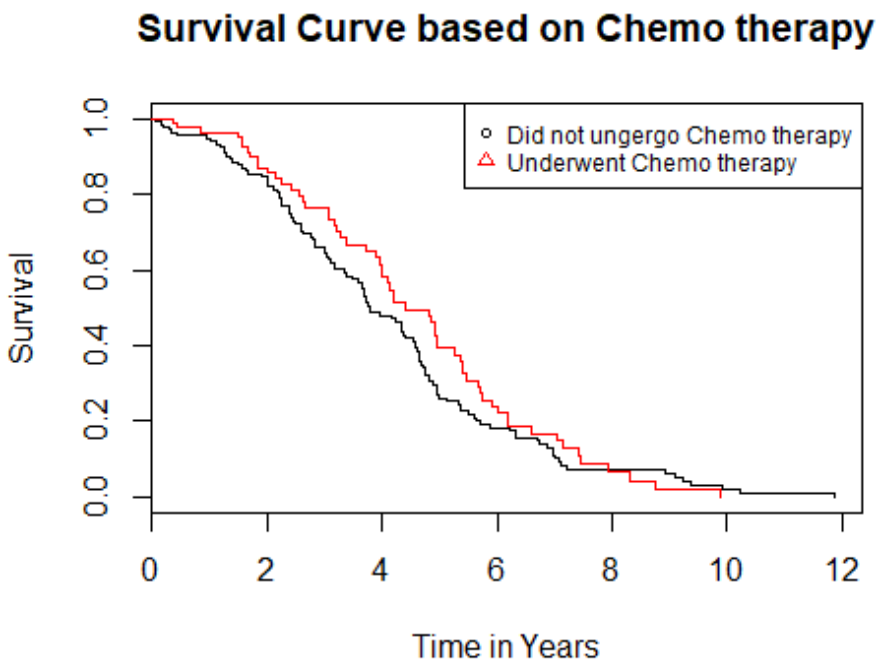
```r
survdiff(test~dukes_stage, data= clinical_data)
```

```
## Call:
## survdiff(formula = test ~ dukes_stage, data = clinical_data)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## dukes_stage=A 41       39     48.4      1.84     2.742
## dukes_stage=B 94       80     65.5      3.20     5.223
## dukes_stage=C 91       57     62.0      0.41     0.646
##
##  Chisq= 5.7  on 2 degrees of freedom, p= 0.0583
```

```r
survfit(test~adjCTX, data = clinical_data)
```

```
## Call: survfit(formula = test ~ adjCTX, data = clinical_data)
##
##             n events median 0.95LCL 0.95UCL
## adjCTX=N 139    116   3.77    3.50    4.59
## adjCTX=Y  87     60   4.41    3.98    5.41
```

```r
plot(survfit(test~adjCTX, data = clinical_data), col = 1:2,xlab = "Time in
Years", ylab ="Survival")
legend(x = "topright",legend=c("Did not ungergo Chemo therapy","Underwent
Chemo therapy"),pch = seq(1,2) ,bty ="o",col=seq(1,2),cex = 0.75)
title("Survival Curve based on Chemo therapy")
```



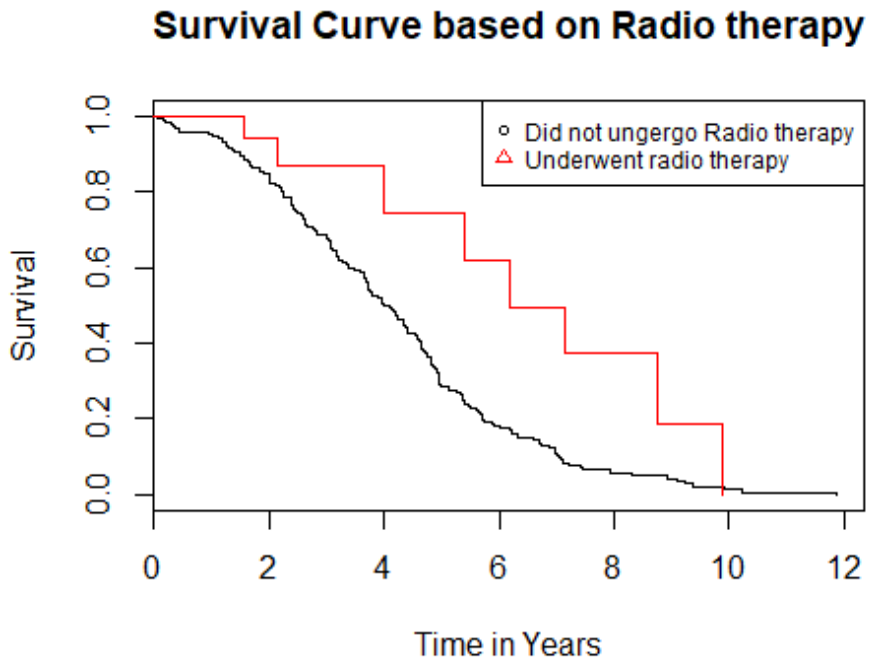
```r
survdiff(test~adjCTX, data = clinical_data)
```

```
## Call:
## survdiff(formula = test ~ adjCTX, data = clinical_data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## adjCTX=N 139      116    109.4     0.400      1.08
## adjCTX=Y  87       60     66.6     0.657      1.08
##
##  Chisq= 1.1  on 1 degrees of freedom, p= 0.299
```

```r
survfit(test~adjXRT, data = clinical_data)
```

```
## Call: survfit(formula = test ~ adjXRT, data = clinical_data)
##
##             n events median 0.95LCL 0.95UCL
## adjXRT=N 204    168   4.09    3.68    4.58
## adjXRT=Y  22      8   6.20    5.41      NA
```

```r
plot(survfit(test~adjXRT, data = clinical_data), col = 1:2,xlab = "Time in
Years", ylab ="Survival")
legend(x = "topright",legend=c("Did not ungergo Radio therapy","Underwent
radio therapy"),pch = seq(1,2) ,bty ="o",col=seq(1,2),cex = 0.75)
title("Survival Curve based on Radio therapy")
```



```r
survdiff(test~adjXRT, data = clinical_data)
```

```
## Call:
## survdiff(formula = test ~ adjXRT, data = clinical_data)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## adjXRT=N 204      168    158.7     0.539       5.6
## adjXRT=Y  22        8     17.3     4.961       5.6
##
##  Chisq= 5.6  on 1 degrees of freedom, p= 0.0179
```

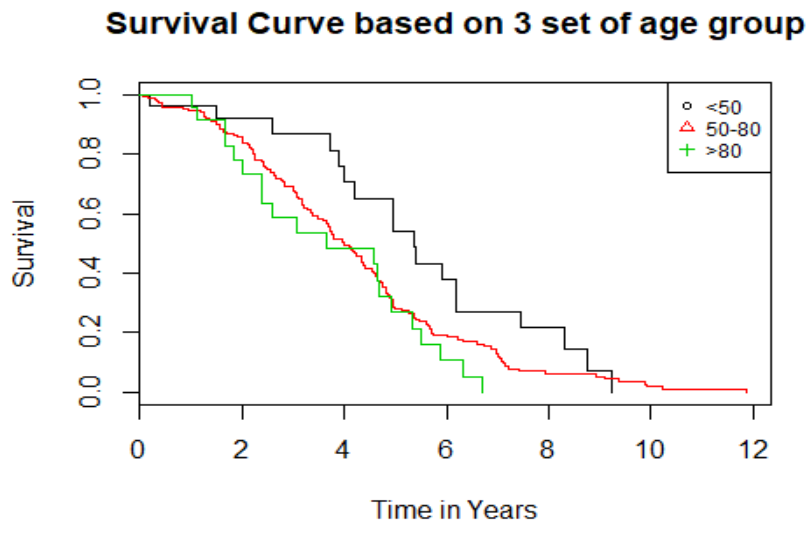
```r
clinical_data$agecat = cut(clinical_data$age_diag,breaks = c(0,50,80,Inf))
table(clinical_data$agecat)
```

```
##
##   (0,50]  (50,80] (80,Inf]
##       27      173       26
```

```r
survfit(test~agecat,data = clinical_data)
```

```
## Call: survfit(formula = test ~ agecat, data = clinical_data)
##
##                    n events median 0.95LCL 0.95UCL
## agecat=(0,50]     27     18   5.38    4.20    8.29
## agecat=(50,80]   173    138   3.99    3.66    4.53
## agecat=(80,Inf]   26     20   3.65    2.40    5.49
```

```r
plot(survfit(test~agecat,data = clinical_data), col = 1:3,xlab = "Time in
Years", ylab = "Survival")
legend(x = "topright",legend=c("<50","50-80",">80"),pch = seq(1,3) ,bty
="o",col=seq(1,3),cex = 0.75)
title("Survival Curve based on 3 set of age group")
```



```
survdiff(test~agecat, data = clinical_data)

## Call:
## survdiff(formula = test ~ agecat, data = clinical_data)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## agecat=(0,50]    27       18     27.4     3.244     3.931
## agecat=(50,80]  173      138    133.8     0.129     0.546
## agecat=(80,Inf]  26       20     14.7     1.897     2.105
##
##  Chisq= 5.4  on 2 degrees of freedom, p= 0.0671
```

We can notice that there are some patients who has taken both the therapy. so I have decided to see the effect of it. In order to do that I have created a new categorical variable named “treatment_type”" with 4 values “No Treatment”, “Chemo”, “Radiation” and “Both”.

From the table we can see that all the patients that has taken the radio therapy has also take the chemo treatment and only one patient had just the radio therapy and even that observation has been censored. So, we have decided to remove this observation from the dataset to make the further analysis easier.

```
clinical_data$treatment_type = "No Treatment"
clinical_data$treatment_type[clinical_data$adjXRT == "Y" ] <- "Radiation"
clinical_data$treatment_type[clinical_data$adjCTX == "Y"] <- "Chemo"
clinical_data$treatment_type[clinical_data$adjXRT == "Y"&
clinical_data$adjCTX == "Y"] <- "Both"

table(clinical_data$treatment_type , dnn = "Kind of treatment underwent by
patients")

## Kind of treatment underwent by patients
##         Both        Chemo No Treatment    Radiation
##           21           66          138            1

table(clinical_data$treatment_type[clinical_data$dfs_event == 1], dnn = "Kind
of treatment underwent by patients and also not being censored")

## Kind of treatment underwent by patients and also not being censored
##         Both        Chemo No Treatment
##            8           52          116
```

New dataset with one observation less than the original data.

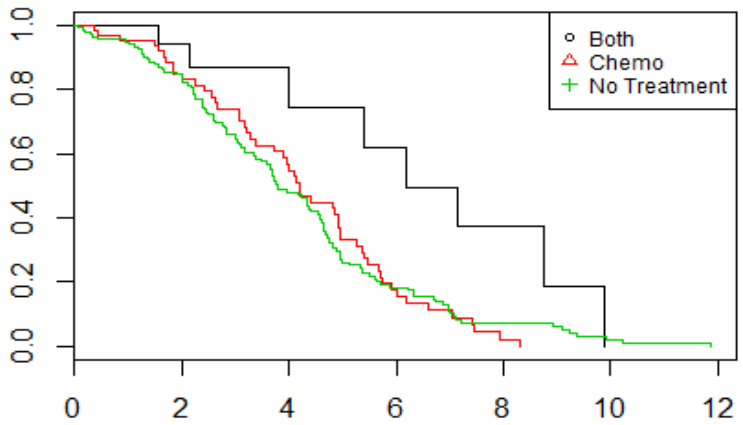
```
clinical_data = subset(clinical_data,!(clinical_data$treatment_type ==
"Radiation"))
```

From the Kaplan-Meier test we can say that the patients who took both treatments has higher survival time than others. So, the significance show by the variable adjXRT is because it indicated the patients who has taken both the therapy.

```
survfit(test~treatment_type, data = clinical_data)

## Call: survfit(formula = test ~ treatment_type, data = clinical_data)
##
##                                n events median 0.95LCL 0.95UCL
## treatment_type=Both           21      8   6.20    5.41      NA
## treatment_type=Chemo          66     52   4.20    3.73    4.96
## treatment_type=No Treatment 138    116   3.77    3.50    4.59

plot(survfit(test~treatment_type, data = clinical_data), col = 1:3,pch =
seq(1,3))
legend(x = "topright",legend=c("Both","Chemo","No Treatment"),pch = seq(1,3)
,bty ="o",col=seq(1,3),cex = 0.75)
```




```
survdiff(test~treatment_type, data =  clinical_data)

## Call:
## survdiff(formula = test ~ treatment_type, data = clinical_data)
##
##                                N Observed Expected (O-E)^2/E (O-E)^2/V
## treatment_type=Both           21        8     17.2     4.928     5.565
## treatment_type=Chemo          66       52     49.4     0.134     0.192
## treatment_type=No Treatment 138      116    109.4     0.402     1.084
##
##  Chisq= 5.6  on 2 degrees of freedom, p= 0.0619
```

But in real world we have cases where the patients can take just radio therapy without taking chemo.so I have decided to create two multivariate models.

model1: with 3 variables dukes_stage, age and treatment_type

model2: with 3 variables dukes_stage, age and adjXRT

From summary of model1 we can see that the variables selected has significance.

Variable: dukes_stage

"dukes_stageA" is taken as the base value. Positive coefficient implies the increase in risk factor which corresponds to decrease in the survival time. so as seen in the survival plot before, survival time of observations with stage A is higher than stage C which is higher than stage B.

variable: treatment_type

"Both"" is taken as the base value. Positive coefficient implies the increase in risk factor which corresponds to decrease in the survival time. so as seen in the survival plot before, survival time of patients who took both the treatment is more than patients who just took chemo and surgery.

Inference we made by performing CoxRegression corresponds to the Survival graph derived from the Kaplan-Meier test (univariate model)

smaller P-value from Wald test and likelihood ratio shows that it is a good model.

```
model1 = coxph(test~dukes_stage + treatment_type + age_diag, data =
clinical_data)

summary(model1)
```

```
## Call:
## coxph(formula = test ~ dukes_stage + treatment_type + age_diag,
##     data = clinical_data)
##
##   n= 225, number of events= 176
##
##                                coef exp(coef) se(coef)     z Pr(>|z|)
## dukes_stageB               0.515506  1.674485 0.208679 2.470   0.0135 *
## dukes_stageC               0.319213  1.376044 0.257185 1.241   0.2145
## treatment_typeChemo        0.881626  2.414824 0.385896 2.285   0.0223 *
## treatment_typeNo Treatment 0.857487  2.357231 0.383139 2.238   0.0252 *
## age_diag                   0.013796  1.013891 0.006722 2.052   0.0401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                            exp(coef) exp(-coef) lower .95 upper .95
## dukes_stageB                   1.674     0.5972    1.1124     2.521
## dukes_stageC                   1.376     0.7267    0.8312     2.278
```

```
## treatment_typeChemo            2.415     0.4141    1.1335     5.145
## treatment_typeNo Treatment     2.357     0.4242    1.1124     4.995
## age_diag                       1.014     0.9863    1.0006     1.027
##
## Concordance= 0.589  (se = 0.026 )
## Rsquare= 0.077   (max possible= 0.999 )
## Likelihood ratio test= 17.94  on 5 df,   p=0.003029
## Wald test            = 16  on 5 df,   p=0.006833
## Score (logrank) test = 16.41  on 5 df,   p=0.005762
```

We can notice that the p-values are high, so we cannot reject the null hypothesis which states that the proportionality of hazard holds.

```
cox.zph(model1)
```

```
##                                 rho   chisq     p
## dukes_stageB                0.01983 0.07272 0.787
## dukes_stageC                0.00298 0.00156 0.968
## treatment_typeChemo        -0.00532 0.00475 0.945
## treatment_typeNo Treatment -0.05719 0.51762 0.472
## age_diag                   -0.01897 0.06770 0.795
## GLOBAL                           NA 2.66679 0.751
```

Creating the model2 but the only difference from model1 is that we have used adjXRT instead of treatment_type variable.

variable: adjXRT (indicates whether the patient has taken radiotherapy or not)

Coefficient is negative which implies the reduction of risk factor and high survival rate.

smaller P-value from Wald test and likelihood ratio shows that it is a good model.

```
model2 <- coxph(test~dukes_stage + adjXRT + age_diag, data = clinical_data)
summary(model2)
```

```
## Call:
## coxph(formula = test ~ dukes_stage + adjXRT + age_diag, data =
clinical_data)
##
##   n= 225, number of events= 176
##
##                   coef exp(coef)  se(coef)      z Pr(>|z|)
## dukes_stageB  0.519569  1.681304  0.205602  2.527   0.0115 *
## dukes_stageC  0.334519  1.397268  0.219844  1.522   0.1281
## adjXRTY      -0.868949  0.419392  0.369903 -2.349   0.0188 *
## age_diag      0.013652  1.013746  0.006604  2.067   0.0387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##              exp(coef) exp(-coef) lower .95 upper .95
## dukes_stageB    1.6813     0.5948    1.1237    2.5157
## dukes_stageC    1.3973     0.7157    0.9081    2.1499
## adjXRTY         0.4194     2.3844    0.2031    0.8659
## age_diag        1.0137     0.9864    1.0007    1.0270
##
## Concordance= 0.59  (se = 0.026 )
## Rsquare= 0.077   (max possible= 0.999 )
## Likelihood ratio test= 17.92  on 4 df,   p=0.001278
## Wald test            = 16.01  on 4 df,   p=0.003003
## Score (logrank) test = 16.41  on 4 df,   p=0.002514
```

We can notice that the p-values are high so we cannot rejec the null hypothesis which states that the propotionality of hazard holds.

```
cox.zph(model2)
```

```
##                  rho chisq     p
## dukes_stageB  0.0357 0.233 0.630
## dukes_stageC  0.0566 0.586 0.444
## adjXRTY       0.0336 0.184 0.668
## age_diag     -0.0345 0.229 0.632
## GLOBAL            NA 1.247 0.870
```

We are using step variable selection method to check whether the model obtained by our inference is same as the one provided by the step function.

```
model1_ss = coxph(test~location + dukes_stage + age_diag + gender +
treatment_type, data = clinical_data)
summary(model1_ss)
```

```
## Call:
## coxph(formula = test ~ location + dukes_stage + age_diag + gender +
##     treatment_type, data = clinical_data)
##
##   n= 225, number of events= 176
##
##                                 coef exp(coef)  se(coef)      z Pr(>|z|)
## locationColon                0.320093  1.377256  0.752512  0.425   0.6706
## locationLeft                -0.108615  0.897076  0.254153 -0.427   0.6691
## locationRight                0.052592  1.053999  0.257069  0.205   0.8379
## dukes_stageB                 0.520345  1.682607  0.211798  2.457   0.0140 *
## dukes_stageC                 0.336841  1.400516  0.258991  1.301   0.1934
## age_diag                     0.011768  1.011837  0.007005  1.680   0.0930 .
## genderM                     -0.064627  0.937417  0.156048 -0.414   0.6788
## treatment_typeChemo          0.881740  2.415097  0.408437  2.159   0.0309 *
## treatment_typeNo Treatment   0.893730  2.444228  0.396828  2.252   0.0243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                              exp(coef) exp(-coef) lower .95 upper .95
## locationColon                   1.3773     0.7261    0.3151     6.019
## locationLeft                    0.8971     1.1147    0.5451     1.476
## locationRight                   1.0540     0.9488    0.6368     1.744
## dukes_stageB                    1.6826     0.5943    1.1110     2.548
## dukes_stageC                    1.4005     0.7140    0.8430     2.327
## age_diag                        1.0118     0.9883    0.9980     1.026
## genderM                         0.9374     1.0668    0.6904     1.273
## treatment_typeChemo             2.4151     0.4141    1.0846     5.378
## treatment_typeNo Treatment      2.4442     0.4091    1.1230     5.320
##
## Concordance= 0.601  (se = 0.026 )
## Rsquare= 0.082   (max possible= 0.999 )
## Likelihood ratio test= 19.31  on 9 df,   p=0.02266
## Wald test            = 17.34  on 9 df,   p=0.04364
## Score (logrank) test = 17.8  on 9 df,   p=0.03761

model1_fit <- step(model1_ss)

## Start:  AIC=1505.07
## test ~ location + dukes_stage + age_diag + gender + treatment_type
##
##                  Df    AIC
## - location        3 1500.1
## - gender          1 1503.2
## <none>              1505.1
## - age_diag        1 1506.0
## - dukes_stage     2 1507.5
## - treatment_type  2 1507.6
##
## Step:  AIC=1500.12
## test ~ dukes_stage + age_diag + gender + treatment_type
##
##                  Df    AIC
## - gender          1 1498.5
## <none>              1500.1
## - dukes_stage     2 1502.2
## - age_diag        1 1502.4
## - treatment_type  2 1503.3
##
## Step:  AIC=1498.45
## test ~ dukes_stage + age_diag + treatment_type
##
##                  Df    AIC
## <none>              1498.5
## - age_diag        1 1500.8
## - dukes_stage     2 1501.0
## - treatment_type  2 1501.5

summary(model1_fit)
```

```
## Call:
## coxph(formula = test ~ dukes_stage + age_diag + treatment_type,
##     data = clinical_data)
##
##   n= 225, number of events= 176
##
##                               coef exp(coef) se(coef)     z Pr(>|z|)
## dukes_stageB               0.515506  1.674485 0.208679 2.470   0.0135 *
## dukes_stageC               0.319213  1.376044 0.257185 1.241   0.2145
## age_diag                   0.013796  1.013891 0.006722 2.052   0.0401 *
## treatment_typeChemo        0.881626  2.414824 0.385896 2.285   0.0223 *
## treatment_typeNo Treatment 0.857487  2.357231 0.383139 2.238   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                            exp(coef) exp(-coef) lower .95 upper .95
## dukes_stageB                   1.674     0.5972    1.1124     2.521
## dukes_stageC                   1.376     0.7267    0.8312     2.278
## age_diag                       1.014     0.9863    1.0006     1.027
## treatment_typeChemo            2.415     0.4141    1.1335     5.145
## treatment_typeNo Treatment     2.357     0.4242    1.1124     4.995
##
## Concordance= 0.589  (se = 0.026 )
## Rsquare= 0.077   (max possible= 0.999 )
## Likelihood ratio test= 17.94  on 5 df,   p=0.003029
## Wald test            = 16  on 5 df,   p=0.006833
## Score (logrank) test = 16.41  on 5 df,   p=0.005762
```

```
model2_ss = coxph(test~location + dukes_stage + age_diag + gender + adjXRT +
adjCTX, data = clinical_data)
summary(model2_ss)
```

```
## Call:
## coxph(formula = test ~ location + dukes_stage + age_diag + gender +
##     adjXRT + adjCTX, data = clinical_data)
##
##   n= 225, number of events= 176
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## locationColon  0.320093  1.377256  0.752512  0.425   0.6706
## locationLeft  -0.108615  0.897076  0.254153 -0.427   0.6691
## locationRight  0.052592  1.053999  0.257069  0.205   0.8379
## dukes_stageB   0.520345  1.682607  0.211798  2.457   0.0140 *
## dukes_stageC   0.336841  1.400516  0.258991  1.301   0.1934
## age_diag       0.011768  1.011837  0.007005  1.680   0.0930 .
## genderM       -0.064627  0.937417  0.156048 -0.414   0.6788
## adjXRTY       -0.881740  0.414062  0.408437 -2.159   0.0309 *
## adjCTXY       -0.011990  0.988082  0.216433 -0.055   0.9558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                exp(coef) exp(-coef) lower .95 upper .95
## locationColon     1.3773     0.7261    0.3151     6.019
## locationLeft      0.8971     1.1147    0.5451     1.476
## locationRight     1.0540     0.9488    0.6368     1.744
## dukes_stageB      1.6826     0.5943    1.1110     2.548
## dukes_stageC      1.4005     0.7140    0.8430     2.327
## age_diag          1.0118     0.9883    0.9980     1.026
## genderM           0.9374     1.0668    0.6904     1.273
## adjXRTY           0.4141     2.4151    0.1860     0.922
## adjCTXY           0.9881     1.0121    0.6465     1.510
##
## Concordance= 0.601  (se = 0.026 )
## Rsquare= 0.082   (max possible= 0.999 )
## Likelihood ratio test= 19.31  on 9 df,   p=0.02266
## Wald test            = 17.34  on 9 df,   p=0.04364
## Score (logrank) test = 17.8  on 9 df,   p=0.03761
```

From the test we can see that step function is giving the same model as the one predicted by us.

```
model2_fit <- step(model2_ss)
```

```
## Start:  AIC=1505.07
## test ~ location + dukes_stage + age_diag + gender + adjXRT +
##     adjCTX
##
##               Df    AIC
## - location     3 1500.1
## - adjCTX       1 1503.1
## - gender       1 1503.2
## <none>           1505.1
## - age_diag     1 1506.0
## - dukes_stage  2 1507.5
## - adjXRT       1 1508.6
##
## Step:  AIC=1500.12
## test ~ dukes_stage + age_diag + gender + adjXRT + adjCTX
##
##               Df    AIC
## - adjCTX       1 1498.1
## - gender       1 1498.5
## <none>           1500.1
## - dukes_stage  2 1502.2
## - age_diag     1 1502.4
## - adjXRT       1 1504.6
##
## Step:  AIC=1498.14
## test ~ dukes_stage + age_diag + gender + adjXRT
##
```

```
##               Df    AIC
## - gender       1 1496.5
## <none>           1498.1
## - dukes_stage  2 1500.5
## - age_diag     1 1500.5
## - adjXRT       1 1503.3
##
## Step:  AIC=1496.46
## test ~ dukes_stage + age_diag + adjXRT
##
##               Df    AIC
## <none>           1496.5
## - age_diag     1 1498.9
## - dukes_stage  2 1499.2
## - adjXRT       1 1501.5
```

```
summary(model2_fit)
```

```
## Call:
## coxph(formula = test ~ dukes_stage + age_diag + adjXRT, data =
clinical_data)
##
##   n= 225, number of events= 176
##
##                   coef exp(coef)  se(coef)      z Pr(>|z|)
## dukes_stageB  0.519569  1.681304  0.205602  2.527   0.0115 *
## dukes_stageC  0.334519  1.397268  0.219844  1.522   0.1281
## age_diag      0.013652  1.013746  0.006604  2.067   0.0387 *
## adjXRTY      -0.868949  0.419392  0.369903 -2.349   0.0188 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## dukes_stageB    1.6813     0.5948    1.1237    2.5157
## dukes_stageC    1.3973     0.7157    0.9081    2.1499
## age_diag        1.0137     0.9864    1.0007    1.0270
## adjXRTY         0.4194     2.3844    0.2031    0.8659
##
## Concordance= 0.59  (se = 0.026 )
## Rsquare= 0.077   (max possible= 0.999 )
## Likelihood ratio test= 17.92  on 4 df,   p=0.001278
## Wald test            = 16.01  on 4 df,   p=0.003003
## Score (logrank) test = 16.41  on 4 df,   p=0.002514
```

But, when we look for the better fitted model using Akaike information criterion, model2 is best fitted than model1.

```
AIC(model1)
```

```
## [1] 1498.447
```

```
AIC(model2)
```

```
## [1] 1496.461
```

Conclusion

While building model2 we have considered the variable adjXRT (indicates whether the patient has taken radiotherapy or not). From the study of our data we have seen that effect of the adjXRT (radiotherapy treatment) is not only based on radiation but based on combined effect of both chemotherapy and radiotherapy. But in real world situation we may have some patients who don't have a combined treatment but just have only one of the treatment (chemotherapy without radiation and vice-versa).

AIC value shows that model2 is better than model1, however from our comparison to real case situation we can conclude that model2 is overfitting the data.

As the result of the above inference we choose to select model1 over model2.