SURVIVAL ANALYSIS OF EXPRESSION DATA FROM PRIMARY COLORECTAL CANCERS

Author: Jasmy Elizabeth Joson

A19 Cohort

Submitted On : May 7, 2020.

1. Description of data

Samples were taken from colorectal cancers in surgically resected specimens in 290 colorectal cancer patients. The data for which the disease free survival time and censoring information was not available were removed. The final number of patients for which the data was available was 226. Survival analysis was done with this final data. Survival refers to the survival of patients from progression of cancer after treatment.

The variables that were available in the data are:

Table 1.1

Variables	Description
location	Tumour location
dukes_stage	Cancer stage (Duke's classification)
age_diag	Age at diagnosis
gender	Gender
dfs_time	Disease Free Survival (DFS) time, in months
dfs_event	DFS event: 1=event time, 0=censoring time
adjXRT	Adjuvant radiation therapy
adjCTX	Adjuvant chemotherapy

```
## Importing libraries
```{r}
library(survival)
library(tidyverse)
library(pROC)
```

Loading and viewing data
```{r}
dataCRC <- load("CRC_226_GSE14333.RData")
clinical_metadata
cdat <- clinical_data
table(cdat$dfs_event)
head(cdat)
```
```

Data Modification

The below columns were added which are modified from the existing ones:

Table 1.2

Variables	Description
dfs_timeYears	Disease Free Survival (DFS) time, in years
age_group2	identify the age group in (25-50) and (above 50+)
age_group3	identify the age group in (25-39), (40-59), (60-79), (80+)
adjT2	shows whether the patient has undergone an adjuvant therapy or not
adjT3	shows whether the patient has undergone: no adjuvant therapy(None), both adjuvant therapy(Both), one of the adjuvant therapy(Single)

Event Distribution

Table 1.3

No. of Censorings (0)	No. of Events (1)
50	176

First few datas:

Table 1.4

sampleID	location	dukes_stage	age_diag	gender	dfs_time	dfs_event	adjXRT	adjCTX	dfs_timeYears	age_group2	age_group3	adjT2	adjT3
GSM358341	Right	A	78.00	M	3.64	1.00	N	N	0.30	50+	70+	N	None
GSM358342	Rectum	A	53.00	F	14.53	0.00	N	N	1.21	50+	49-69	N	None
GSM358343	Left	A	80.00	F	16.47	1.00	N	N	1.38	50+	70+	N	None
GSM358344	Left	A	58.00	M	19.75	1.00	N	N	1.65	50+	49-69	N	None
GSM358345	Left	A	81.00	M	20.02	1.00	N	N	1.67	50+	70+	N	None
GSM358346	Right	A	57.00	M	23.96	1.00	N	N	2.00	50+	49-69	N	None

2. Basic descriptive statistics

Summary of data

```
cdat <- mutate(cdat, dfs_timeYears = dfs_time * 30.5 / 365.25,
               age_group2 = as.factor(ifelse(age_diag <= 50, '25-50', '50+')),
               age_group3 = as.factor(ifelse(age_diag <=48, '25-48',
                                             ifelse(age_diag > 48 & age_diag <= 69, '49-69', '70+'))),
               adjT2 = as.factor(ifelse(adjXRT =='N' & adjCTX == 'N', 'N', 'Y')),
               adjT3 = as.factor(ifelse(adjXRT =='N' & adjCTX == 'N', 'None',
                                        ifelse(adjXRT =='Y' & adjCTX == 'Y', 'Both','Single'))))
summary(cdat)
```

Table 2.1

sampleID	location	dukes_stage	age_diag	gender	dfs_time	dfs_event	adjXRT	adjCTX	dfs_timeYears
Length:226	Rectum: 30	A:41	Min. :26.00	F:106	Min. : 0.92	Min. :0.0000	N:204	N:139	Min. : 0.07682
Class :character	Colon : 2	B:94	1st Qu.:58.00	M:120	1st Qu.: 22.28	1st Qu.:1.0000	Y: 22	Y: 87	1st Qu.: 1.86069
Mode :character	Left : 93	C:91	Median :67.00		Median : 38.46	Median :1.0000			Median : 3.21158
	Right :101		Mean :66.03		Mean : 43.52	Mean :0.7788			Mean : 3.63396
			3rd Qu.:75.00		3rd Qu.: 59.50	3rd Qu.:1.0000			3rd Qu.: 4.96852
			Max. :92.00		Max. :142.55	Max. :1.0000			Max. :11.90356

age_group2	age_group3	adjT2	adjT3
25-50: 27	25-48: 22	N:138	Both : 21
50+ :199	49-69:103	Y: 88	None :138
	70+ :101		Single: 67

From the summary of the data we find,

- The sample size is 226(sampleID).
- Tumour have 4 locations in this sample and has data of the patients at cancer stage A, B and C (location and dukes_stage).
- Youngest age at which cancer diagnosed in the sample is 26 and oldest is 92(age_diag).
- There are 106 females and 120 males in the data(gender).
- The trial period is 142.55 months or 11.9 years(dfs_time and dfs_timeYears).
- Out of 226 patients, 22 have undergone adjuvant radiation therapy and 87 have undergone adjuvant chemo therapy(adjXRT and adjCTX).
- There are 27 patients in the age group 25 to 50 and 199 above 50(age_group2).
- 22 patients between age 25 and 48. Almost equal number of patients in 49 – 69 and above 70(age_group3).
- 88 patients have undergone atleast one of the adjuvant therapies(adjT2).
- 21 has undergone both adjuvant therapies, 67 has undergone exactly one of the adjuvant therapies and 138 have not undergone any of the adjuvant therapies(adjT3).

3. Questions asked, methods used, results

The data has the detail whether the patients were undergoing adjuvant therapies. So I am analysing if the adjuvant therapies makes any difference in progression of cancer based on cancer stage, age, gender and tumour location of the subject.
Note: For the analysis, I am considering time in year.
Finally, the created model is used to find the high risk subjects and for prediction.

Method 1 : Non Parametric Analysis for single sample

Survival function with Kaplan-Meier Estimator

```
Call: survfit(formula = Surv(dfs_timeYears, dfs_event) ~ 1, data = cdat)

      n  events  median 0.95LCL 0.95UCL
 226.00  176.00    4.16    3.74    4.65
```

```
plot(fitKM, main = "Kaplan-Meier estimator", conf.int = T,
     ylab = "Survival probability", xlab = "time (years)")
abline(h=0.5, col = "blue")
```

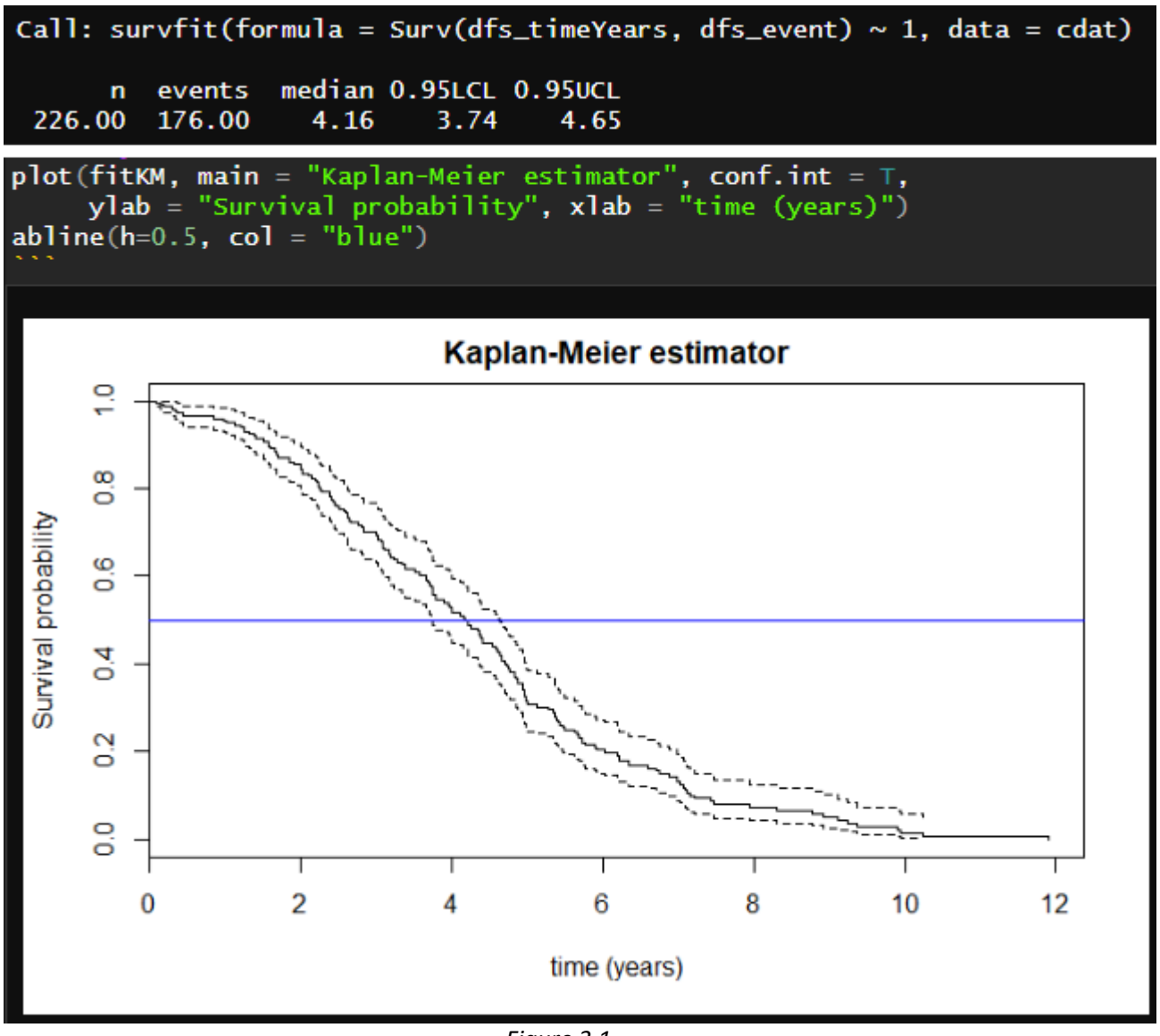


Figure 3.1

- According to Kaplan-Meier test, there is a 50% chance that cancer might make a progress within 4.16 years (49.8 months) with a confidence interval of 3.74 - 4.65 (44.8 - 55.7).

Nelson-Aalen Estimator

```
Call: survfit(formula = Surv(dfs_timeYears, dfs_event) ~ 1, data = cdat,
    type = "fh")

      n  events  median 0.95LCL 0.95UCL
 226.00 176.00    4.21    3.74    4.65
```

```
plot(fitNA, main = "Nelson-Aalen estimator",
     ylab = "Survival probability", xlab = "time (years)")
abline(h=0.5, col = "blue")
```

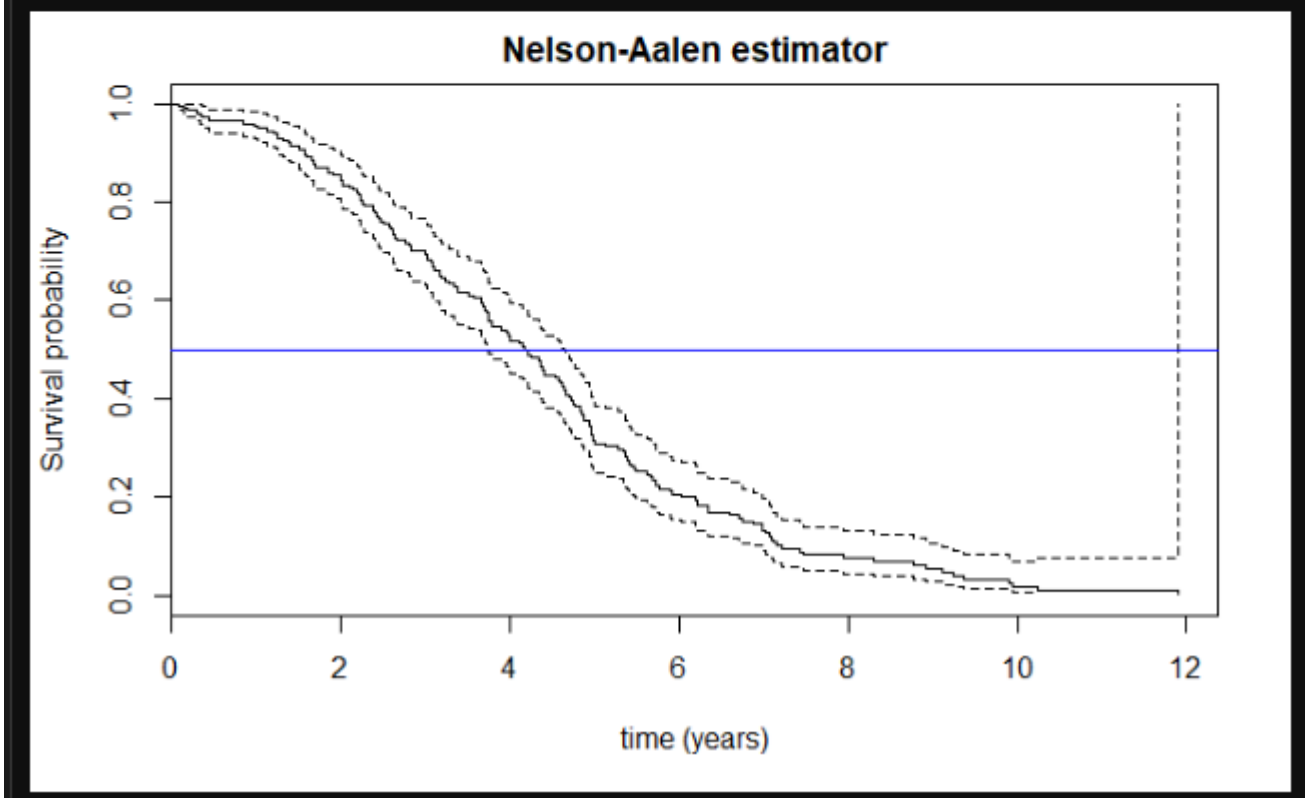


Figure 3.2

- With Nelson-Aalen estimator, there is a 50% chance of progression of cancer within 4.21 years (slightly different from Kaplan-Meier), but confidence interval exactly same as Kaplan-Meier 3.74 - 4.65.

Method 2 : Non Parametric Analysis (group based)

Here the LogRank test is used to do the statistical test based on : adjuvant therapies (adjXRT, adjCTX), age_group2 and gender. LogRank works for more than 2 groups also, but since it is hard to read we do this only for 2 group covariates.

LogRank based on gender

```
Call:
survdiff(formula = Surv(dfs_timeYears, dfs_event) ~ gender,
data = cdat)

            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= 0.4
```

Figure 3.3

From the LogRank test we can see:

- p-value = 0.4 which is very large. So we do not reject H0 i.e. there is no significant difference between female and male survival. Since there is no difference between both groups, it is better to consider this as a single sample to get a single estimate rather than stratifying them as female and male.

LogRank based on age_group2

```
Call:
survdiff(formula = Surv(dfs_timeYears, dfs_event) ~ age_group2,
    data = cdat)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
age_group2=25-50  27       18     27.4     3.244      3.93
age_group2=50+   199      158    148.6     0.599      3.93

 Chisq= 3.9  on 1 degrees of freedom, p= 0.05
```

```
Call: survfit(formula = Surv(dfs_timeYears, dfs_event) ~ age_group2,
    data = cdat)

                 n events median 0.95LCL 0.95UCL
age_group2=25-50  27     18   5.39    4.21    8.31
age_group2=50+   199    158   4.00    3.66    4.54
```

```
plot(fitKMage2, main = "Kaplan-Meier estimator", col = 1:2,
     ylab = "Survival probability", xlab = "time (years)")
abline(h=0.5, col = "blue")
legend(8.5, 1, legend=c("Age 25-50", "Age 50+"), col=1:2, lty=1:1)
```

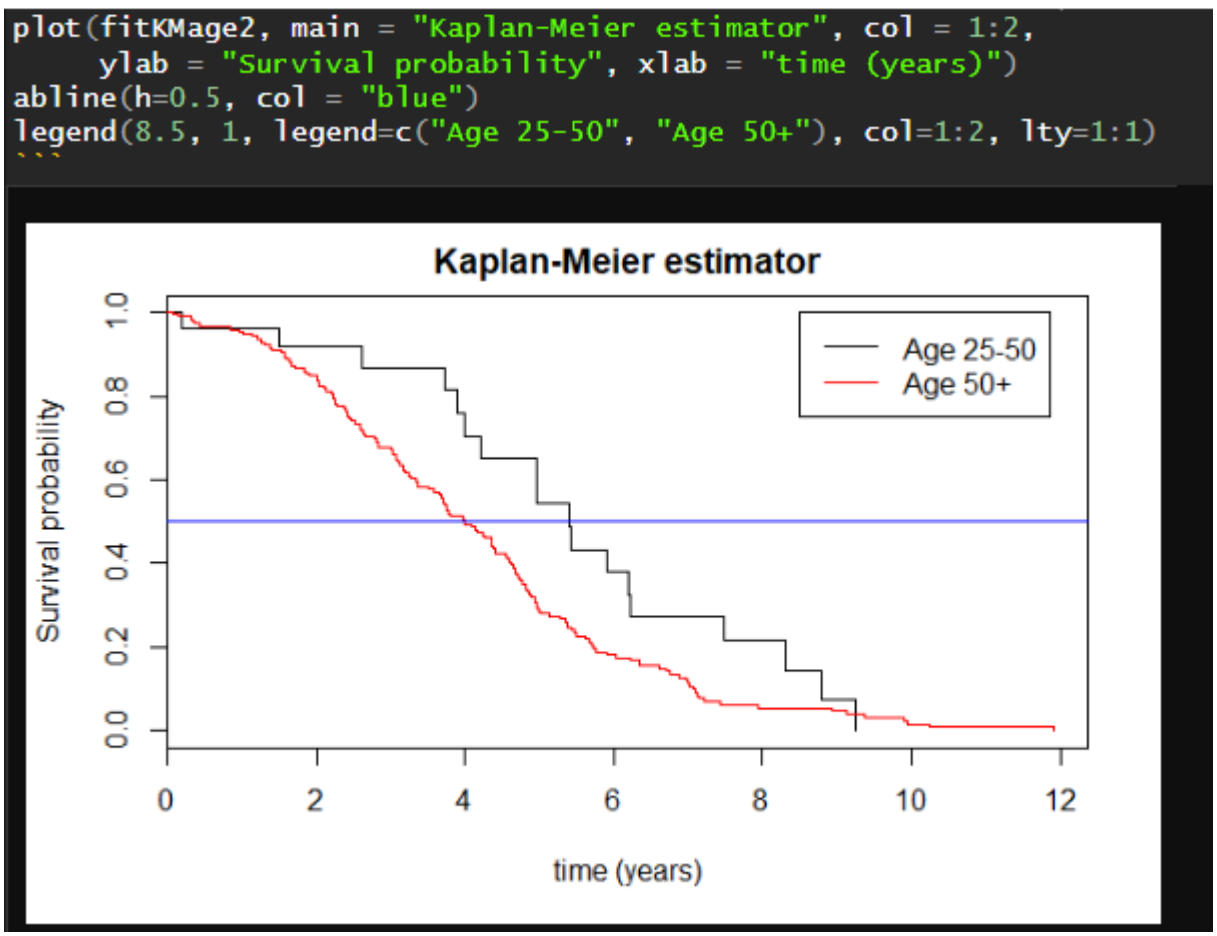


Figure 3.4

The null hypothesis for the test is
$H0 : S_a(t) = S_b(t)$ where
$S_a(t)$ is the survival function of patients in age group 25 – 50,
$S_b(t)$ is the survival function of patients of age greater than 50

- From LogRank, p-value = 0.05 which is small and so we do not reject H1. i.e. the survival function of these 2 groups are different. So we do the survival function (Kaplan-Meier) for this samples.

From Kaplan-Meier we can infer that:

- Patients below and equal to 50 years has higher survival compared to those above 50.
- There is a 50% chance of cancer progression within 5.39 years with a confidence interval of 4.21 to 8.31 for patients in age group 25-50.
- There is a 50% chance of cancer progression within 4 years with a confidence interval of 3.66 to 4.54 for patients above 50 years.

But each age groups have different stages of cancer which might confound our results. So to check this, a stratified LogRank test is done based on dukes_stage.

Stratified LogRank for age_group2 based on dukes_stage

```
Call:
survdiff(formula = Surv(dfs_timeYears, dfs_event) ~ age_group2 +
    strata(dukes_stage), data = cdat)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
age_group2=25-50 27       18     28.5     3.883       5.1
age_group2=50+   199      158    147.5     0.751       5.1

 Chisq= 5.1  on 1 degrees of freedom, p= 0.02
```

Figure 3.5

- The p-value = 0.02 which is still very less. So there is significant difference in survival between the patients for age groups 25-50 and 50+ irrespective of the cancer stage.

LogRank based on adjuvant chemotherapy (adjCTX)

```
Call:
survdiff(formula = Surv(dfs_timeYears, dfs_event) ~ adjCTX,
data = cdat)

           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= 0.3
```

```
Call: survfit(formula = Surv(dfs_timeYears, dfs_event) ~ adjCTX,
data = cdat)

           n events median 0.95LCL 0.95UCL
adjCTX=N 139    116   3.78    3.51    4.60
adjCTX=Y  87     60   4.41    3.99    5.42
```

Figure 3.6

- From LogRank, p-value = 0.3 which is very large. So we do not reject H0 i.e.. there is no significant difference in survival function between whether undergone adjCTX or not. Since there is no significant difference between both groups, it is better to consider this as a single sample to get a single estimate rather than stratifying them based on adjuvant chemo therapy.
- From the Kaplan-Meier, the median survival for those undergone chemo therapy and not are 4.41 and 3.78 respectively but with an overlapping confidence interval of (3.99-5.42) and (3.51-4.60). So it is not significant.

LogRank based on adjuvant radiation therapy (adjXRT)

```
Call:
survdiff(formula = Surv(dfs_timeYears, dfs_event) ~ adjXRT, data
= cdat)

           N Observed Expected (O-E)^2/E (O-E)^2/V
adjXRT=Y  22        8     17.3     4.961       5.6
adjXRT=N 204      168    158.7     0.539       5.6

 Chisq= 5.6  on 1 degrees of freedom, p= 0.02
```

```
Call: survfit(formula = Surv(dfs_timeYears, dfs_event) ~ adjXRT,
data = cdat)

           n events median 0.95LCL 0.95UCL
adjXRT=N 204    168   4.10    3.69    4.58
adjXRT=Y  22      8   6.21    5.42      NA
```

- From LogRank test, p-value = 0.02 which is small and so we do not reject H1. i.e. the survival function of these 2 groups are different. So we will check the Kaplan-Meier estimates(survival function) for both.

From the Kaplan-Meier test we can infer:

- There is a 50% chance of cancer progression within 4.1 years for patients who have not undergone adjuvant radiation therapy.
- There is a 50% chance of cancer progression within 6.21 years for patients who have undergone adjuvant radiation therapy.

```
plot(fitKMadjXRT, main = "Kaplan-Meier estimator", col = 1:2,
conf.int=T, ylab = "Survival probability", xlab = "time (years)")
abline(h=0.5, col = "blue")
legend(8, 0.85, legend=c("No RT", "Undergone RT"), col=1:2,
lty=1:1)
```

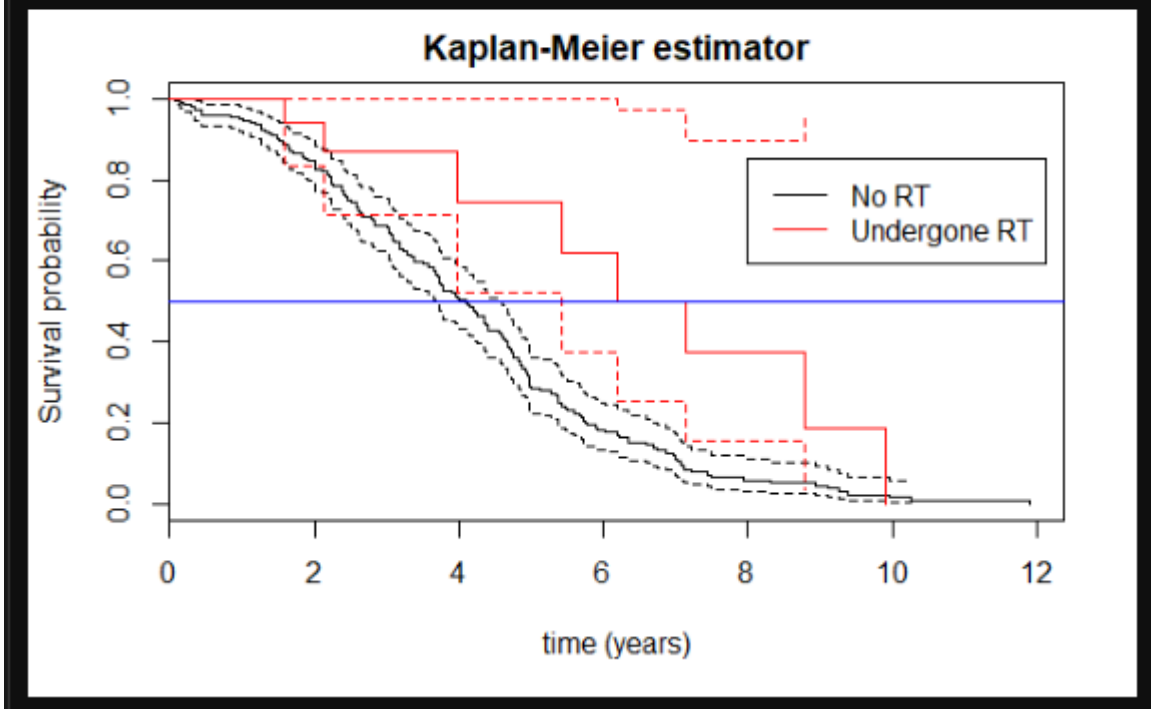


Figure 3.7

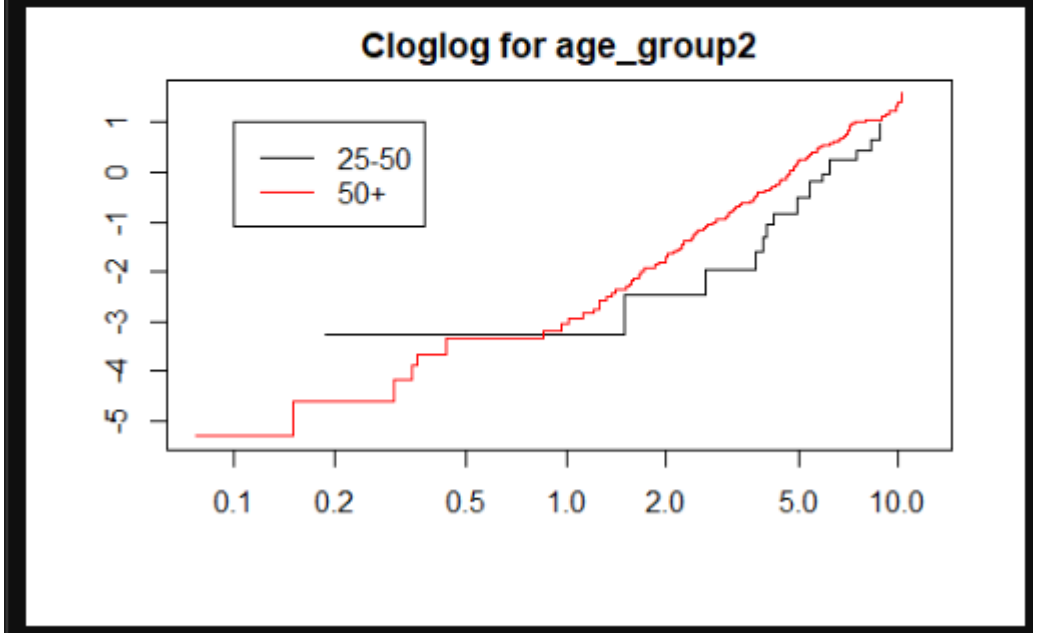
- So undergoing adjuvant radiation therapy is statistically better than not undergoing one.
- In the survival curve we can see that patients undergone radiation therapy has a higher survival.

Conclusion from LogRank:

From logrank tests above we clearly see that:

- There is no significant difference in survival based on gender and adjuvant chemotherapy.
- There is difference in survival of patients based on age_group2 and adjuvant radiation therapy. But we will confirm this with Cloglog plot of survival.

```
clogAge2 <- survfit(Surv(dfs_timeYears, dfs_event) ~
age_group2, data = cdat)
plot(clogAge2, main = "Cloglog for age_group2", fun=
"cloglog", col = 1:2)
legend(0.1, 1, legend=c("25-50", "50+"), col=1:2, lty=1:1)
```



From the survival curves in cloglog scale:

- For age_group2, the curves are not parallel, so the risks are not proportional and the model does not hold.

```
clogXRT <- survfit(Surv(dfs_timeYears, dfs_event) ~ adjXRT,
data = cdat)
plot(clogXRT, main = "Cloglog for adjXRT", fun= "cloglog",
col = 1:2)
legend(0.1, 1, legend=c("No adjXRT", "Undergone adjXRT"),
col=1:2, lty=1:1)
```

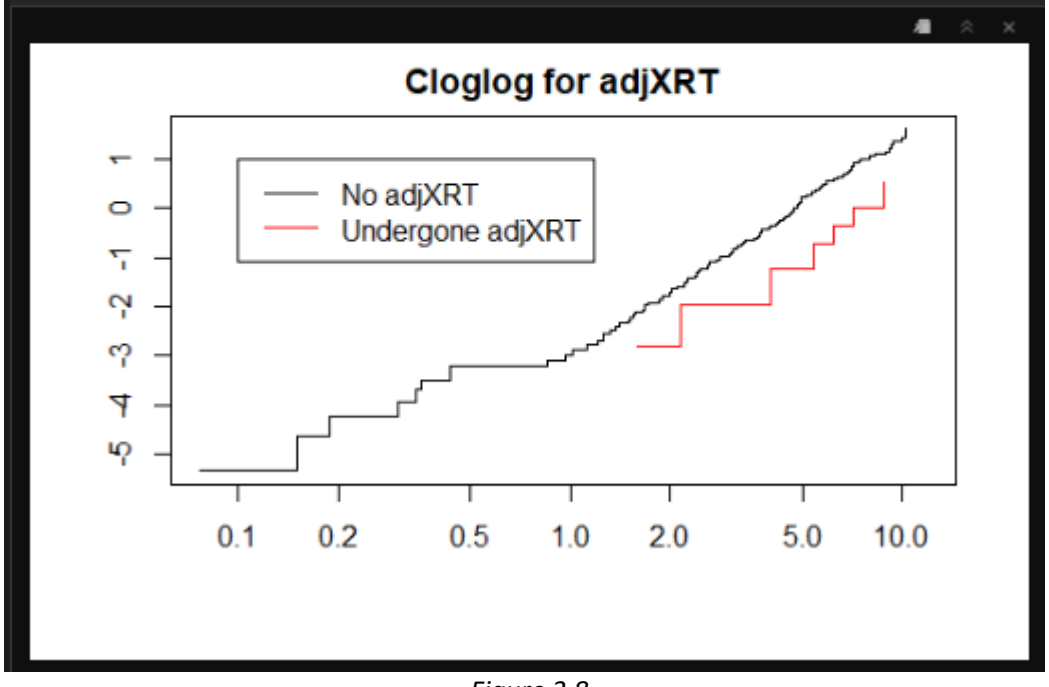


Figure 3.8

From the survival curves in cloglog scale:

- For adjXRT, the curves are parallel, so this is good. But the data points are less.

Method 3 : Regression : Cox Proportional Hazards Model (coxph)

LogRank works for more than 2 groups also, but hard to read. So Coxph is used to analyse continuous variates and those with more than 2 groups. Here I check for survival dependency on age at diagnosis, location of tumour, stage of cancer, age_group3, adjT3.

Cox Regression based on age_diag

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ age_diag,
data = cdat)

  n= 226, number of events= 176

             coef exp(coef) se(coef)     z Pr(>|z|)
age_diag 0.015151  1.015266 0.006517 2.325   0.0201 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
age_diag     1.015      0.985     1.002     1.028

Concordance= 0.563  (se = 0.026 )
Rsquare= 0.025   (max possible= 0.999 )
Likelihood ratio test= 5.63  on 1 df,   p=0.02
Wald test            = 5.41  on 1 df,   p=0.02
Score (logrank) test = 5.42  on 1 df,   p=0.02
```

Figure 3.9

- Age at which cancer is diagnosed is significant as p-value is 0.02 (less than 5%). The positive coefficient means higher the age, higher the risk. The hazards ratio 1.015 indicates an increase in risk of 1.5% as the age increase by one year.

Cox Regression based on age_group3

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ age_group3,
    data = cdat)

  n= 226, number of events= 176

                  coef exp(coef) se(coef)     z Pr(>|z|)
age_group349-69 0.4736    1.6057   0.2853 1.660   0.0970 .
age_group370+   0.5198    1.6817   0.2826 1.839   0.0659 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
age_group349-69     1.606     0.6228    0.9179     2.809
age_group370+       1.682     0.5946    0.9665     2.926

Concordance= 0.539  (se = 0.024 )
Rsquare= 0.017   (max possible= 0.999 )
Likelihood ratio test= 3.89  on 2 df,   p=0.1
Wald test            = 3.42  on 2 df,   p=0.2
Score (logrank) test = 3.49  on 2 df,   p=0.2
```

Figure 3.10

- Beta coefficient is positive for the age groups '49-69' and '70+', they have a higher risk compared to the age group '25-48'. The hazards ratio is 1.60 and 1.68 for the age groups '49-69' and '70+' respectively, indicating 60 and 68 percent higher risk. But since the p-value greater than 5%, this is not very significant.

Cox Regression based on adjT3

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ adjT3, data = cdat)

  n= 226, number of events= 176

              coef exp(coef) se(coef)     z Pr(>|z|)
adjT3None   0.8354    2.3058   0.3671 2.276   0.0228 *
adjT3Single 0.8288    2.2905   0.3825 2.167   0.0302 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

            exp(coef) exp(-coef) lower .95 upper .95
adjT3None       2.306     0.4337     1.123     4.735
adjT3Single     2.291     0.4366     1.082     4.847

Concordance= 0.545  (se = 0.022 )
Rsquare= 0.03   (max possible= 0.999 )
Likelihood ratio test= 6.79  on 2 df,   p=0.03
Wald test            = 5.26  on 2 df,   p=0.07
Score (logrank) test = 5.56  on 2 df,   p=0.06
```

Figure 3.11

- Risk is twice higher for patients who have not done any of the adjuvant therapies(adjT3None) compared to those who have done both adjuvant therapies. Hazards ratio is 2.3 with confidence interval of 1.12-4.74 and the p-value = 0.02 makes this very significant.
- Risk is twice higher for patients who have done only one of the adjuvant therapies(adjT3Single) compared to those who have done both adjuvant therapies. Hazards ratio is 2.3 with confidence interval of 1.08-4.85 and the p-value = 0.03 makes this very significant.

Cox Regression based on dukes_stage

```
cdat <- mutate(cdat,
               dukes_stage = relevel(dukes_stage, ref = "B"))
fit_stageCPH <- coxph(Surv(dfs_timeYears, dfs_event) ~ dukes_stage,
                      data = cdat)
summary(fit_stageCPH)
```

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ dukes_stage,
    data = cdat)

  n= 226, number of events= 176

                coef exp(coef) se(coef)      z Pr(>|z|)
dukes_stageA -0.4433    0.6419   0.2033 -2.180   0.0292 *
dukes_stageC -0.2870    0.7505   0.1740 -1.649   0.0991 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
dukes_stageA    0.6419      1.558    0.4309    0.9562
dukes_stageC    0.7505      1.332    0.5337    1.0556

Concordance= 0.543  (se = 0.024 )
Rsquare= 0.025   (max possible= 0.999 )
Likelihood ratio test= 5.64  on 2 df,   p=0.06
Wald test            = 5.62  on 2 df,   p=0.06
Score (logrank) test = 5.69  on 2 df,   p=0.06
```

Figure 3.12

- Stage A has significantly lower risk compared to stage B with p-value=0.02.
- Stage C is not significantly different from stage B as p value=0.099 which is greater than 5%.

From above coxph models, we find age_diag, adjT3 and dukes_stage are significant. Now coxph with multiple covariates are done.

Cox Regression with multiple covariates

Coxph model with age_diag, dukes_stage, location, adjT3

```
cdat <- mutate(cdat,
               dukes_stage = relevel(dukes_stage, ref = "B"))
fit_multiCPH1 <- coxph(Surv(dfs_timeYears, dfs_event) ~ dukes_stage
                       + age_diag + adjT3 + location, data = cdat)
summary(fit_multiCPH1)
```

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ dukes_stage +
    age_diag + adjT3 + location, data = cdat)

  n= 226, number of events= 176

                 coef exp(coef)  se(coef)      z Pr(>|z|)
dukes_stageA  -0.530301  0.588428  0.210442 -2.520   0.0117 *
dukes_stageC  -0.186661  0.829725  0.200659 -0.930   0.3522
age_diag       0.011761  1.011830  0.007016  1.676   0.0937 .
adjT3None      0.885983  2.425368  0.395475  2.240   0.0251 *
adjT3Single    0.870571  2.388274  0.405962  2.144   0.0320 *
locationColon  0.362292  1.436619  0.745457  0.486   0.6270
locationLeft  -0.105455  0.899915  0.253375 -0.416   0.6773
locationRight  0.065156  1.067326  0.254606  0.256   0.7980
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
dukes_stageA     0.5884     1.6994    0.3896    0.8888
dukes_stageC     0.8297     1.2052    0.5599    1.2295
age_diag         1.0118     0.9883    0.9980    1.0258
adjT3None        2.4254     0.4123    1.1172    5.2651
adjT3Single      2.3883     0.4187    1.0778    5.2922
locationColon    1.4366     0.6961    0.3333    6.1926
locationLeft     0.8999     1.1112    0.5477    1.4787
locationRight    1.0673     0.9369    0.6480    1.7580

Concordance= 0.598  (se = 0.025 )
Rsquare= 0.081   (max possible= 0.999 )
Likelihood ratio test= 19.17  on 8 df,   p=0.01
Wald test            = 17.24  on 8 df,   p=0.03
Score (logrank) test = 17.7  on 8 df,   p=0.02
```

Figure 3.13

- Negative coefficient for dukes_stageA with a hazards ratio of 0.59 indicates a 41% lower risk for stage A cancer compared to stage B cancer with confidence interval 0.39-0.89. The p-value=0.01 makes this highly significant.
- Negative coefficient for dukes_stageC means a lower risk in stage C than B. But since the p-value=0.35 is very high this is not significant. ie Stage C cancer is not significantly different from stage B.
- The positive coefficient for the age means higher the age, higher the risk. The hazards ratio 1.012 indicates an increase in risk of 1.2% as the age increase by one year. But this is not very significant as p-value=0.09 is greater than 5%.
- Positive coefficients for adjT3None and adjT3Single means lower risk for patients who did both adjuvant radiation and chemo therapies compared to others. Hazards ratio is 2 in both cases which means risk is twice higher compared to a patient undergone both therapies with confidence intervals of 1.12-5.27 and 1.08-5.29. This is significant as p-value is 0.03 which is less than 5%.
- Beta coefficient for all tumour locations compared to 'Colon' is negative means lower risk in all locations compared to 'Colon', but p-value for all 3 is very large (0.63, 0.52, 0.68) which makes this insignificant.

Since location is not very significant, we will check for the model without location : **age_diag, dukes_stage, adjT3**

Coxph model with age_diag, dukes_stage, adjT3

```
cdat <- mutate(cdat, dukes_stage = relevel(dukes_stage, ref = "B"))
fit_multiCPH2 <- coxph(Surv(dfs_timeYears, dfs_event) ~ dukes_stage
                       + age_diag + adjT3, data = cdat)
summary(fit_multiCPH2)
```

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ dukes_stage +
    age_diag + adjT3, data = cdat)

  n= 226, number of events= 176

                coef exp(coef)  se(coef)      z Pr(>|z|)
dukes_stageA -0.515634  0.597122  0.208678 -2.471   0.0135 *
dukes_stageC -0.196653  0.821476  0.201246 -0.977   0.3285
age_diag      0.013825  1.013921  0.006721  2.057   0.0397 *
adjT3None     0.857250  2.356672  0.383138  2.237   0.0253 *
adjT3Single   0.881035  2.413396  0.385907  2.283   0.0224 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
dukes_stageA     0.5971     1.6747    0.3967    0.8989
dukes_stageC     0.8215     1.2173    0.5537    1.2187
age_diag         1.0139     0.9863    1.0007    1.0274
adjT3None        2.3567     0.4243    1.1122    4.9938
adjT3Single      2.4134     0.4144    1.1328    5.1418

Concordance= 0.589  (se = 0.026 )
Rsquare= 0.076   (max possible= 0.999 )
Likelihood ratio test= 17.96  on 5 df,   p=0.003
Wald test            = 16.03  on 5 df,   p=0.007
Score (logrank) test = 16.43  on 5 df,   p=0.006
```

Figure 3.14

- Positive coefficient for the age means higher the age, higher the risk. The hazards ratio 1.014 with a confidence interval of 1.00-1.03 indicates an increase in risk of 1.4% as the age increase by one year. This is significant as p-value=0.039 is less than 5%.

Also adjT3 is a combination variable of adjXRT and adjCTX. Since adjCTX is not significant, we will create the model with adjXRT instead of adjT3.

Coxph model with age_diag, dukes_stage, adjXRT

```
cdat <- mutate(cdat, dukes_stage = relevel(dukes_stage, ref = "B"),
               adjXRT = relevel(adjXRT, ref = "Y"))
fit_multiCPH3 <- coxph(Surv(dfs_timeYears, dfs_event) ~ dukes_stage +
                         age_diag + adjXRT, data = cdat)
summary(fit_multiCPH3)
```

```
Call:
coxph(formula = Surv(dfs_timeYears, dfs_event) ~ dukes_stage +
    age_diag + adjXRT, data = cdat)

  n= 226, number of events= 176

                coef exp(coef)  se(coef)      z Pr(>|z|)
dukes_stageA -0.519712  0.594692  0.205600 -2.528   0.0115 *
dukes_stageC -0.185204  0.830934  0.175863 -1.053   0.2923
age_diag      0.013664  1.013758  0.006604  2.069   0.0385 *
adjXRTN       0.870799  2.388819  0.369904  2.354   0.0186 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
dukes_stageA    0.5947     1.6815    0.3975    0.8898
dukes_stageC    0.8309     1.2035    0.5887    1.1729
age_diag        1.0138     0.9864    1.0007    1.0270
adjXRTN         2.3888     0.4186    1.1570    4.9323

Concordance= 0.59  (se = 0.026 )
Rsquare= 0.076   (max possible= 0.999 )
Likelihood ratio test= 17.98  on 4 df,   p=0.001
Wald test            = 16.06  on 4 df,   p=0.003
Score (logrank) test = 16.46  on 4 df,   p=0.002
```

Figure 3.15

- Positive coefficient for adjXRTN means higher risk for patients who have not done adjuvant radiation therapy compared to others. Hazards ratio is 2.4 with confidence interval of 1.16-4.93 which means risk is twice higher compared to the patients undergone adjuvant radiation therapy. This is very significant as p-value is 0.019 which is less than 5%.

Comparing the models with concordance scores

Comparing the above 2 models : **dukes_stage + age_diag + adjT3** & **dukes_stage + age_diag + adjXRT** with concordance scores(*Figure 3.14 & Figure 3.15*), the second one seems to be the better one with higher concordance score (0.59).

Comparing the models with AIC scores

```
fits <- list(fit_multiCPH1 = fit_multiCPH1,
             fit_multiCPH2 = fit_multiCPH2,
             fit_multiCPH3 = fit_multiCPH3)
sapply(fits, AIC)
```

```
fit_multiCPH1 fit_multiCPH2 fit_multiCPH3
     1503.307      1498.518      1496.490
```

Figure 3.16

- On comparing the above models, we get **dukes_stage + age_diag + adjXRT** as the better model with lower AIC score of 1496.490.

Note:
fit_multiCPH1 => dukes_stage + age_diag + adjT3 + location
fit_multiCPH2 => dukes_stage + age_diag + adjT3
fit_multiCPH2 => dukes_stage + age_diag + adjXRT

Automatic model selection based on AIC (stepwise AIC)

```
fitCPHFull <- coxph(Surv(dfs_timeYears, dfs_event) ~ location +
dukes_stage + age_diag + gender + adjXRT + adjCTX, data = cdat)
fitCPHInitial <- coxph(Surv(dfs_timeYears, dfs_event) ~ 1, data =
cdat)
MAICmodel <- step(fitCPHInitial, direction = "forward", steps = 15,
scope = list(lower = fitCPHInitial, upper = fitCPHFull))
```

```
Start:  AIC=1506.48
Surv(dfs_timeYears, dfs_event) ~ 1

              Df    AIC
+ adjXRT       1 1501.6
+ age_diag     1 1502.8
+ dukes_stage  2 1504.8
<none>           1506.5
+ adjCTX       1 1507.4
+ gender       1 1507.7
+ location     3 1509.1

Step:  AIC=1501.63
Surv(dfs_timeYears, dfs_event) ~ adjXRT

              Df    AIC
+ dukes_stage  2 1498.9
+ age_diag     1 1499.2
<none>           1501.6
+ gender       1 1502.8
+ adjCTX       1 1503.6
+ location     3 1505.1

Step:  AIC=1498.92
Surv(dfs_timeYears, dfs_event) ~ adjXRT + dukes_stage

           Df    AIC
+ age_diag  1 1496.5
<none>        1498.9
+ gender    1 1500.5
+ adjCTX    1 1500.8
+ location  3 1502.4

Step:  AIC=1496.49
Surv(dfs_timeYears, dfs_event) ~ adjXRT + dukes_stage + age_diag

           Df    AIC
<none>        1496.5
+ gender    1 1498.2
+ adjCTX    1 1498.5
+ location  3 1501.3
```

Figure 3.17

This resulted in the same previous model as **dukes_stage + age_diag + adjXRT**.

Model Diagnostics

To see if the model is good enough, **Shoenfeld residuals** is also checked.

```
res.scho <- cox.zph(coxph(Surv(dfs_timeYears, dfs_event) ~ dukes_stage
                          + age_diag + adjXRT, data = cdat))
res.scho
```

```
                rho chisq     p
dukes_stageA -0.0356 0.232 0.630
dukes_stageC  0.0299 0.156 0.693
age_diag     -0.0347 0.231 0.631
adjXRTN      -0.0343 0.192 0.661
GLOBAL            NA 1.261 0.868
```

Figure 3.18

The p-values are large, so the model is good. So the final model chosen is with covariates : **dukes_stage, age_diag and adjXRT**.

4. Results and Conclusions

From the model chosen (age_diag+adjXRT+dukes_stage (*Figure 3.15*)) and other tests above, we get the following results:

- Undergoing adjuvant radiation therapy is statistically better than not undergoing one.
- There is a 50% chance of cancer progression within 4.1 years for patients who have not undergone adjuvant radiation therapy.
- There is a 50% chance of cancer progression within 6.21 years for patients who have undergone adjuvant radiation therapy.
- The patients who have not undergone adjuvant radiation therapy has twice higher risk compared to those who have done adjuvant radiation therapy. Hazards ratio is 2.4 with confidence interval of 1.16-4.93. This is very significant as p-value is 0.019 which is less than 5%.
- Risk increases with age of the subject.
- Higher the age, higher the risk. The hazards ratio 1.014 with a confidence interval of 1.00-1.03 indicates an increase in risk of 1.4% as the age increase by one year. The p-value=0.04 makes this significant.
- Stage B and C subjects have higher risk and they are not significantly different from each other.
- There is 41% lower risk for stage A cancer compared to stage B cancer with confidence interval 0.39-0.89. The p-value=0.01 makes this highly significant.
- Progression of cancer is not related to location of the tumour.
- There is no significant difference between female and male survival.
- There is no significant difference in survival function based on whether the patient has undergone adjuvant chemotherapy or not.

Top High Risk subjects

Based on the model chosen, we will now identify the subjects at high risk:

```
d_new <- select(cdat, -dfs_time, -dfs_event, -dfs_timeYears)
d_segmented <-
  d_new %>%
  mutate(risk_score = predict(fit_multiCPH3, newdata = d_new, type = "lp"))
d_segmented %>%
  arrange(desc(risk_score)) %>%
  head(10)
```

	sampleID	location	dukes_stage	age_diag	gender	adjXRT	adjCTX	age_group2	age_group3	adjT2	adjT3	risk_score
1	GSM358442	Right	B	92.00	F	N	N	50+	70+	N	None	0.61
2	GSM358448	Right	B	92.00	M	N	N	50+	70+	N	None	0.61
3	GSM358408	Left	B	89.00	F	N	N	50+	70+	N	None	0.57
4	GSM358410	Right	B	86.00	F	N	N	50+	70+	N	None	0.53
5	GSM358434	Right	B	86.00	F	N	N	50+	70+	N	None	0.53
6	GSM358407	Right	B	84.00	F	N	N	50+	70+	N	None	0.50
7	GSM358389	Rectum	B	83.00	M	N	N	50+	70+	N	None	0.49
8	GSM358431	Left	B	83.00	F	N	N	50+	70+	N	None	0.49
9	GSM358444	Rectum	B	82.00	M	N	N	50+	70+	N	None	0.47
10	GSM358446	Right	B	82.00	M	N	N	50+	70+	N	None	0.47

Table 4.1

Prediction

To do prediction with the model, the data is split into train and test data. 80% of the data is selected randomly as the train data and the model is trained with this data. We use this model to predict the scores for the test data, find the AUC and plot the ROC. AUC with this model is 0.84.

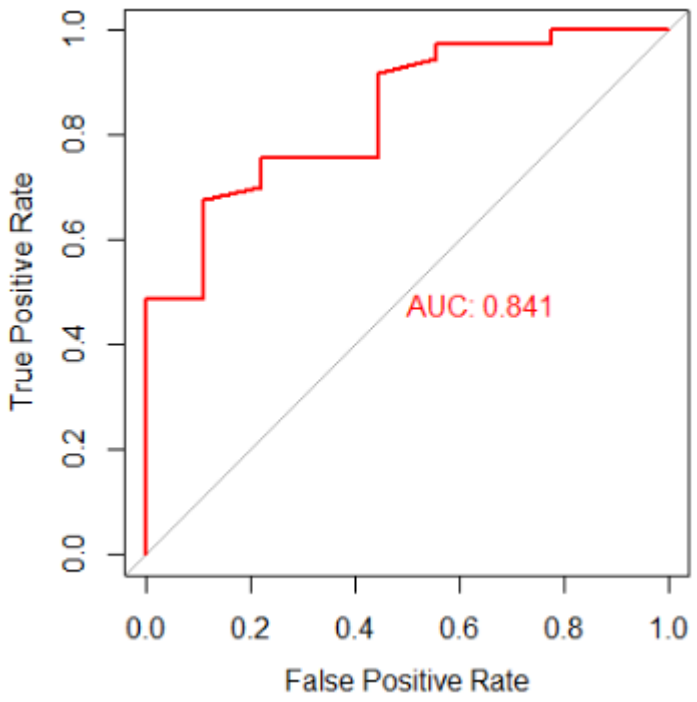


Figure 4.2

```
Call:
roc.default(response = d_predict$dfs_event, predictor =
d_predict$risk_score,     plot = TRUE, legacy.axes = TRUE, xlab = "False
Positive Rate",     ylab = "True Positive Rate", col = 2, print.auc = TRUE)

Data: d_predict$risk_score in 9 controls (d_predict$dfs_event 0) < 37 cases
(d_predict$dfs_event 1).
Area under the curve: 0.8408
```

```
set.seed(6)
len_data = dim(cdat)[1]
index = sample(seq_len(nrow(cdat)), size = len_data*0.8)
train = cdat[index,]
test = cdat[-index,]
modelCPH <- coxph(Surv(dfs_timeYears, dfs_event) ~ dukes_stage  + age_diag + adjXRT, data = train)

d_predict <- test %>%
  mutate(risk_score = predict(modelCPH, newdata = test, type = "lp"))

par(pty = "s")
roc(d_predict$dfs_event, d_predict$risk_score, plot = TRUE, legacy.axes = TRUE,
    xlab = "False Positive Rate", ylab = "True Positive Rate", col = 2, print.auc = TRUE)
```

Prediction of a particular subject

We can also predict the survival of a particular subject using the survfit function as follows:

```
predcph <- survfit(modelCPH, newdata = test[1,], type = "aalen")
summary(predcph)
plot(predcph, ylab = "predicted survival")
```

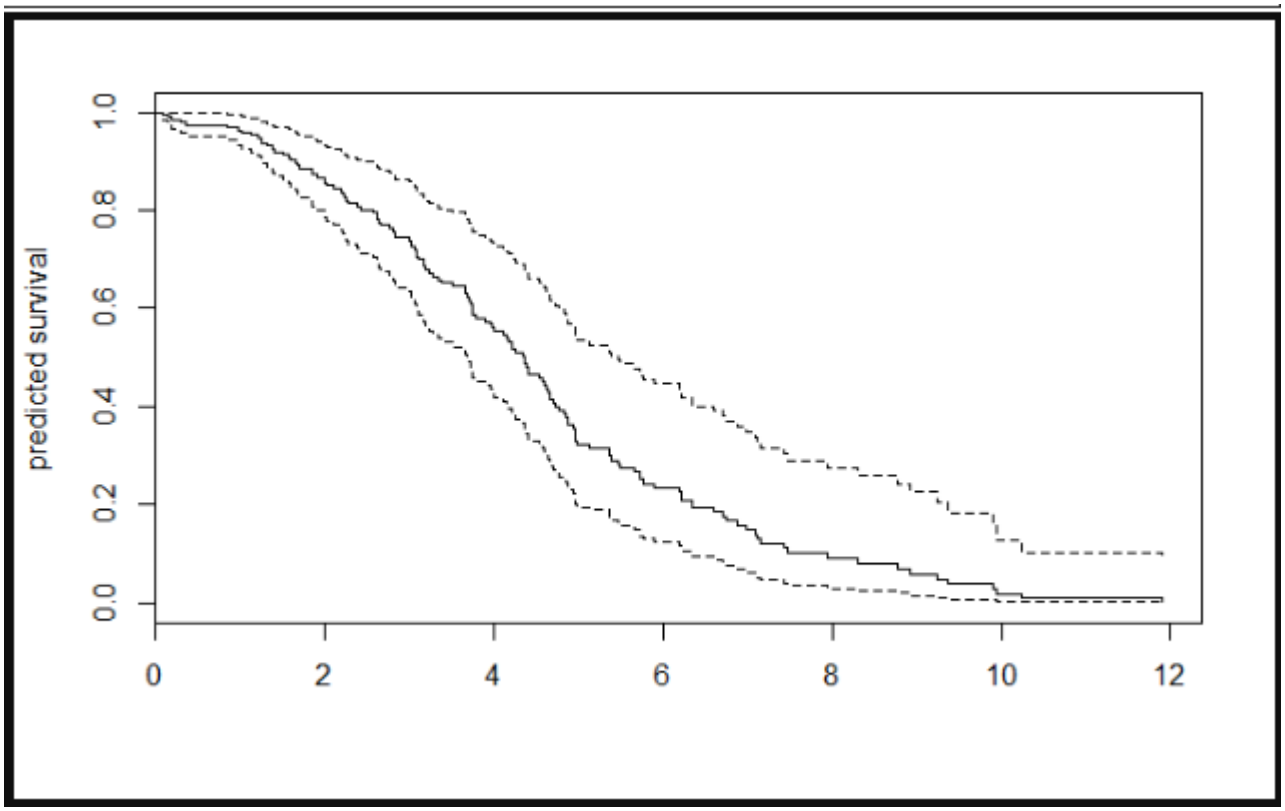


Figure 4.3

```
Call: survfit(formula = modelCPH, newdata = test[1, ], type = "aalen")

    time n.risk n.event survival std.err lower 95% CI upper 95% CI
  0.0768    180       1  0.99494 0.00516      0.98488        1.0000
  0.1503    179       1  0.98987 0.00742      0.97544        1.0000
  0.1887    178       1  0.98481 0.00924      0.96687        1.0000
  0.3040    176       1  0.97972 0.01086      0.95867        1.0000
  0.3541    175       1  0.97462 0.01234      0.95073        0.9991
  0.8517    166       1  0.96928 0.01383      0.94254        0.9968
  0.9687    165       1  0.96394 0.01525      0.93452        0.9943
  1.0104    163       1  0.95858 0.01660      0.92659        0.9917
```