Survival Analysis on Bladder Cancer Patients

Phillip Kim, Allison Tam, Jenny Gao

11/30/2018

Introduction

Our data set contains bladder cancer data. The event of interest for this data set is the survival times for 86 patients until recurrence of bladder cancer tumors occurs. In this data set we are given the outcome variable or the survival time for each patient (inttime), the corresponding censoring status (event), and several other variables. The variable TX is the exposure variable of interest, representing the effect of the drug treatment of thiotepa. The variables NUM and SIZE are control variables of interest, as they are possible confounders of the survival time variable.

The 9 variables included in our data are:

-ID: subject ID

-Event: 0 = censored, 1 = recurrence of cancer found (event of failure is recurrence of cancer)

-Interval: number of data lines for each subject (each event triggers additional line)

-Inttime: length of observation period

-Start: month in which observation starts (unless the data is left-censored)

-Stop: month in which observation ends (unless the data is right-censored)

-Tx: different treatment groups (2 groups), one group is treated with thiotepa (tx=1) and one group is the control group (tx=0)

-Num: initial number of tumors (covariate)

-Size: initial size of the tumor (cm) (covariate)

The scientific questions we will answer in this project are: Does the drug treatment of thiotepa have a significant effect on the survival time for bladder cancer patients? Does the initial number of tumors or the initial size of tumors affect the survival time for bladder cancer patients? Does the survivability distribution of a non-recurrent data set differ from that when considering recurrence?

First, we will import the bladder dataset.

```
library("survival")
bladder <- read.table("bladder.txt", skip=12, header=TRUE)
```

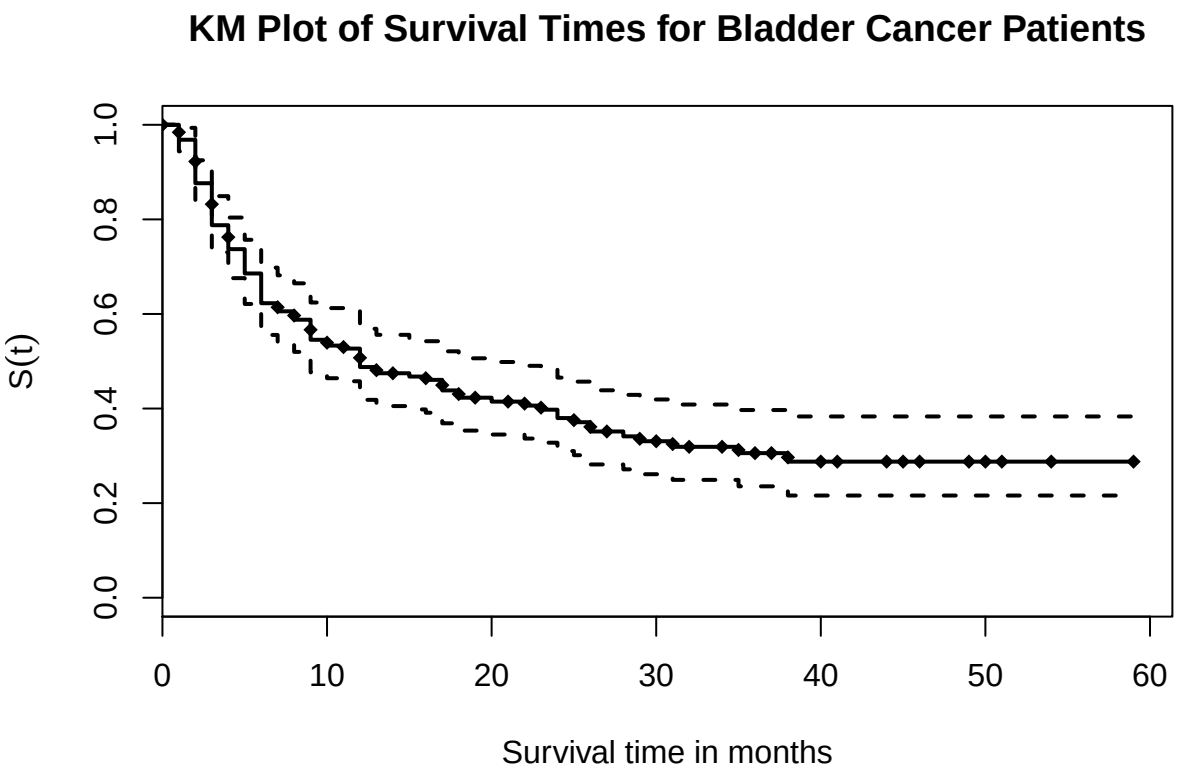
The first part of this report will assume that there is nonrecurrent data; in other words, we assume that no patient has multiple episodes of bladder cancel. The second part of this report, however, will make use of the recurrent data to output a different type of analysis.

Part 1: Nonrecurrent Data

Kaplan Meier Estimate

We first start by estimating and graphing the survivability of the patients using the Kaplan-Meier (KM) method.

```
kmfit = survfit(Surv(INTTIME, EVENT)~1, data=bladder)
plot(kmfit, main="KM Plot of Survival Times for Bladder Cancer Patients", xlab="Survival time in months"
     ylab=expression(hat(S)(t)),lwd=2,mark.time = TRUE,mark=18)
```



From the plot above,we can see that within the first year of treatment, about half of the patients experience recurrence of bladder cancer; however, as time diverges, we see that the survivability probability converges to about 0.3.

To demonstrate this sudden decrease in the survivability probability of the patients, we make use of quantile estimation:

```
quantile(kmfit, probs = c(.1, .3, .5), conf.int = F)
```

```
## 10 30 50
##  2  5 12
```

The quantile function confirms that the median survival time is 12 months, that the patient's probability of not experiencing recurrence of cancer is 90% after 2 months, and that the the patient's probability of not experiencing recurrence of cancer is 70% after 5 months. From the results of the quantile function, there seems to be a sudden drop in the survivability of the patients as indicated from the KM plot.

We want to see the graphical representation of the logarithmic survivability distribution. We do so by making use of Greenwood's formula: $Var(log(S(t_k))) = \sum_{j=1}^{k} \frac{m_j}{n_j(n_j - m_j)}$, where m_j is the number of events and n_j is the number of observations at risk.

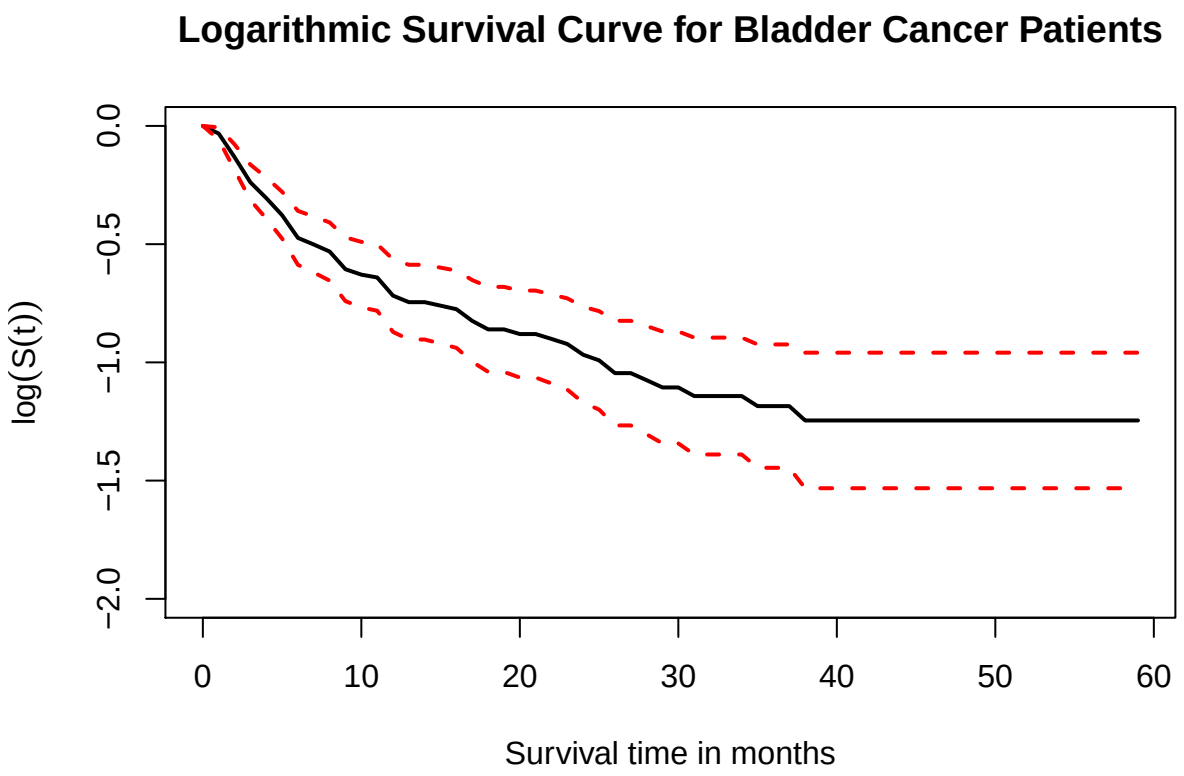
```
mj = kmfit$n.event
nj = kmfit$n.risk

Vj = mj/nj/(nj - mj)
cVj = cumsum(Vj)

lowerCI = log(kmfit$surv) - 1.96*sqrt(cVj)
upperCI = log(kmfit$surv) + 1.96*sqrt(cVj)

plot(kmfit$time,log(kmfit$surv),lwd=2,type="l",ylim=c(-2,0),
xlab="Survival time in months",ylab=expression(log(hat(S)(t))),
main = "Logarithmic Survival Curve for Bladder Cancer Patients" )

lines(kmfit$time,lowerCI,lty=2,col=2,lwd=2)
lines(kmfit$time,upperCI,lty=2,col=2,lwd=2)
```



Comparing the two treatment groups:

From what we have mentioned above, the data set includes data with two treatment groups: one group being treated with thiotepa (TX = 1), and the other being a control group (TX = 0). We want to compare the two groups and see whether the survivability functions for the two groups are similar to each other to determine if thiotepa is an effective treatment.

```
bladtrt = survfit(Surv(INTTIME, EVENT)~TX, data = bladder)
bladtrt
```

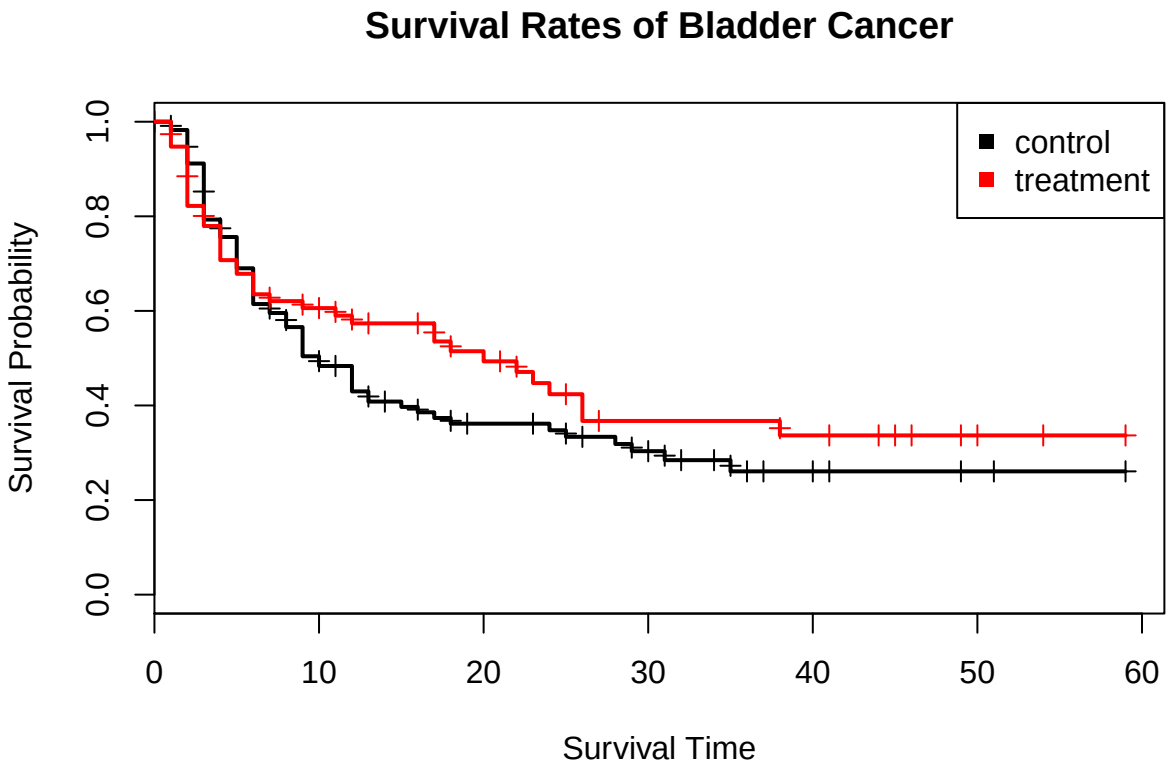
```
## Call: survfit(formula = Surv(INTTIME, EVENT) ~ TX, data = bladder)
##
##        n events median 0.95LCL 0.95UCL
```

```
## TX=0 115     72     10       8      16
## TX=1  76     40     20      11      NA
```

From the above, we can see that the median survival time point for the control group is 10 months, with the control group survivability time having a 95% confidence interval of [8, 16].

We also observe that the median survival time point for the thiotepa group is 20 months, with the thiotepa group having a lower bound of 11 for the survivability time and a missing value for the upper bound.

```
plot(bladtrt, lwd = 2, col = 1:2, mark.time = TRUE, main = "Survival Rates of Bladder Cancer",
     xlab = "Survival Time", ylab = "Survival Probability")
legend("topright", legend = c("control", "treatment"), pch = 15, col = 1:2)
```



Overall, the treatment group seems to have better survivability than the control group; however, we want to test and see if there is significant difference between the two treatments.

Log Rank Test

We conduct log-rank test to test the null hypothesis that the control group is the same as the treatment group:

```
survdiff(Surv(INTTIME, EVENT)~TX, data=bladder)
```

```
## Call:
## survdiff(formula = Surv(INTTIME, EVENT) ~ TX, data = bladder)
##
##        N Observed Expected (O-E)^2/E (O-E)^2/V
## TX=0 115       72     66.6     0.439      1.16
## TX=1  76       40     45.4     0.643      1.16
##
##  Chisq= 1.2  on 1 degrees of freedom, p= 0.3
```

Since our p-value is 0.3, which is greater than $\alpha = 0.05$, we can reject our null hypothesis and conclude that there is a notable difference in the survival curves for the control group and the treatment group.

Model Fitting

Cox PH Regression: Preparing for Fowards Selection

We want to create survival models by relating time of survivability to the covariates of our bladder dataset. The main covariates that we will be analyzing are the NUM, SIZE, and TX variables. We ignore the INTERVAL covariate because this covariate directly works with recurrence of data.

```
fit0 <- coxph(Surv(INTTIME, EVENT)~NUM+SIZE+TX, data=bladder)
anova(fit0)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(INTTIME, EVENT)
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -522.20
## NUM  -519.25 5.8978  1    0.01516 *
## SIZE -519.24 0.0208  1    0.88519
## TX   -517.98 2.5218  1    0.11228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above results, we can see which covariates are most significant based on the p-values. We observe that the NUM covariate is the most significant, then TX, and finally SIZE. This allows us to effectively create other potential models for Forwards Selection down below.

AIC

We can verify our above model by performing forwards selection with AIC, in which we obtain the AIC's for the various models and pick the model with the minimal AIC.

We would first want to create the other models based on the signficance of the covariates that we had analyzed above.

We first have fit1 to be a model which contains NUM and TX covariates:

```
fit1 <- coxph(Surv(INTTIME, EVENT)~NUM+TX, data=bladder)
anova(fit1)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(INTTIME, EVENT)
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -522.20
## NUM  -519.25 5.8978  1    0.01516 *
## TX   -518.00 2.4992  1    0.11390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above result, we can still see that NUM is still the most significant covariate.

Then we have fit2 to be a model which contains the NUM covariate.

```
fit2 <- coxph(Surv(INTTIME, EVENT)~NUM, data=bladder)
anova(fit2)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(INTTIME, EVENT)
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -522.20
## NUM  -519.25 5.8978  1    0.01516 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see from above that the NUM covariate has p-value $= 0.01516 < \alpha = 0.05$.

```
AIC(fit0) #num+size+tx
```

```
## [1] 1041.964
```

```
AIC(fit1) #num+tx
```

```
## [1] 1040.007
```

```
AIC(fit2) #num
```

```
## [1] 1040.506
```

The AIC value for the model, fit1, with the NUM and TX covariates, is 1040.007, which is the smallest AIC value. This reinforces our conclusion that NUM is the only significant covariate.

Model Fitting with Coefficients (Without Recurrence)

The model fitting that has been done thus far does not take into consideration recurrence of bladder cancer, and instead, treats each patient uniquely. We still present the hazard ratios for both the NUM and TX covariates in our model, fit1.

```
fit1
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ NUM + TX, data = bladder)
##
##        coef exp(coef) se(coef)     z      p
## NUM  0.1385    1.1486   0.0492  2.82 0.0048
## TX  -0.3171    0.7283   0.2032 -1.56 0.1186
##
## Likelihood ratio test=8.4  on 2 df, p=0.02
## n= 191, number of events= 112
```

We have from above, $\beta_{\text{NUM}} = 0.1385$, and the 95% confidence interval for β_{NUM} is: $[0.042068, 0.234932]$. The point estimate for the hazard ratio for the NUM covariate is: $e^{0.1385} = 1.14855$, and the 95% confidence interval for the hazard ratio is: $[1.04297, 1.26482]$. From this, we can see that for every unit increase in the NUM covariate, the hazard increases, and so a patient is less likely to survive for longer periods of time.

We also have from above, $\beta_{\text{TX}} = -0.3171$, and the 95% confidence interval for β_{TX} is: $[-0.715372, 0.081172]$. The point estimate for the hazard ratio for the TX covariate is: $e^{-0.3171} = 0.72826$, and the 95% confidence interval for the hazard ratio is: $[0.48901, 1.08456]$. From the point estimate of the hazard ratio for the TX covariate, we can see that for every unit increase of the TX covariate, the hazard decreases, and so a patient is more likely to survive for longer periods of time.

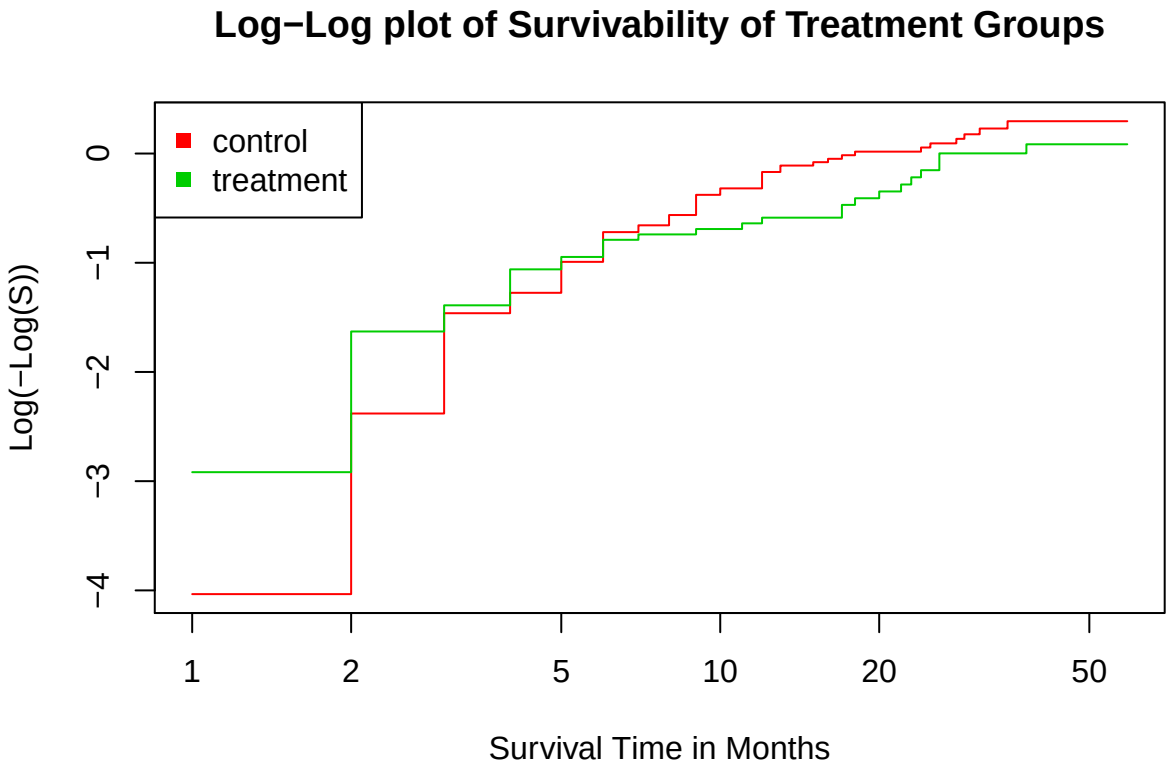
Comparison of Treatments (TX) part 2

We come back to the comparison of the two treatment groups. We will first make use of the log-log plot to determine whether the Cox PH assumption holds for modelling the effects of the treatment groups.

Log-Log Plot

```
split.fit <- survfit(Surv(INTTIME, EVENT)~TX, data=bladder)
plot(split.fit, fun="cloglog", col=c(2,3), xlab="Survival Time in Months", ylab="Log(-Log(S))",
     main = "Log-Log plot of Survivability of Treatment Groups")

legend("topleft", legend = c("control", "treatment"), pch = 15, col = 2:3)
```



For the proportional hazards model to be considered appropriate, the vertical distance between the two curves should be constant. The relationship between log(-logS(t)) and the covariate x is not linear and there is a nonconstant distance in log(-logS(t)).Because the distance between the curves is not constant over time and the two lines intersect at multiple points, we conclude that the Cox PH assumption is not valid for modelling the effects of the treatment groups (TX variable).

However, we decide to test the Cox PH assumption on the treatment covariate and see if our observations from the log-log plot contradict the results from applying the Cox PH assumption to the treatment groups.

Cox PH for Treatments

```
blad = coxph(Surv(INTTIME, EVENT)~TX, data = bladder)
blad
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ TX, data = bladder)
##
##      coef exp(coef) se(coef)     z    p
## TX -0.213     0.808    0.198 -1.08 0.28
##
## Likelihood ratio test=1.18  on 1 df, p=0.3
## n= 191, number of events= 112
```

From the above coxph function, we can estimate the hazard proportion between the control and treatment group: $(h_0(t)e^{\beta x})/h_0(t) = e^{\beta X_1}$. From the above result, we obtain $e^{\beta} = 0.808$, which is the value of our proportion. For one unit increase in X_1, the hazard increases by a factor of e^{β}, for $e^{\beta} < 1$. The risk of recurrence decreases by 19.2% with each unit increase in X_1 (the beta corresponding to TX is negative, so the hazard ratio decreases with time).

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

```
##        2.5 %   97.5 %
## TX 0.5480867 1.191193
```

The 95% confidence interval for the hazards ratio for treatment is (0.5481, 1.1912).

We now use the cox.zph function to perform a test to see if the model is significantly divergent from the proportional hazards model

```
cox.zph(fit1,global = FALSE) #NUM
```

```
##         rho chisq     p
## NUM  0.0679 0.455 0.500
## TX  -0.1307 1.808 0.179
```

Our p-values are both greater than $\alpha = 0.05$, so there is insufficient evidence to suggest that the PH assumption does not hold. This is consistent with our results from the graphs and previous tests. We conclude that the Cox PH is reasonable because there is not significant evidence for us to reject that assumption.

Residuals

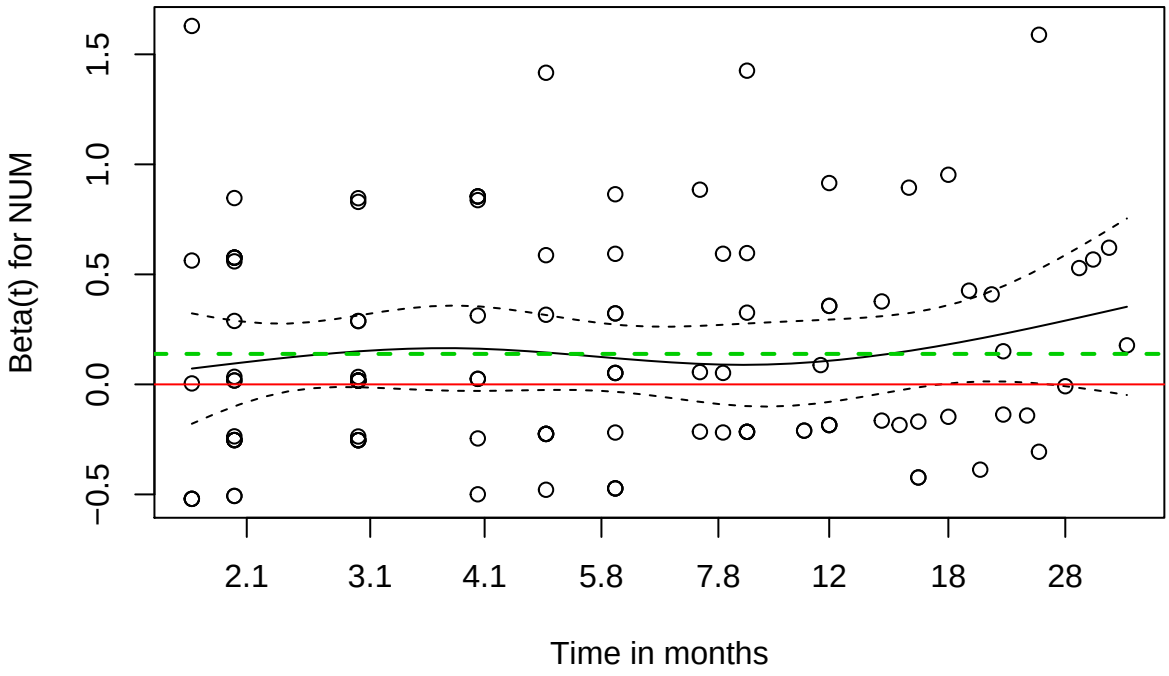
We want to know which covariates of the bladder data should enter the model as independent of time. We have alreayd determined that the NUM covariate is the only covariate such that it has a significant effect on survivability time. We will then look at the fit2 model which contains the covariate NUM.

```
zp = cox.zph(fit1)
zp
```

```
##            rho chisq     p
## NUM     0.0679 0.455 0.500
## TX     -0.1307 1.808 0.179
## GLOBAL      NA 1.952 0.377
```

Our p-values are both greater than $\alpha = 0.05$, which indicates that the Schoenfeld residuals are not constant over time and that the covariate NUM is confirmed to be time independent. We then plot the Schoenfeld residuals below:

```r
plot(zp[1], xlab = "Time in months") # a plot for the 3rd variable in the fit
abline(0,0,col=2)
abline(h=fit1$coefficients[1], col=3, lwd=2, lty=2)
```



From the above plot of the Schoenfeld residuals, we conclude that because the trend for beta versus time is horizontal for the covariate, the PH assumption is true.

EXTENSION

One other covariate which we would like to see whether it fits into our Cox PH regression model is the following: NUM*TX. We again perform backwards elimination to see whether this interaction term has a significant effect on time or not while making use of the control convariates NUM, SIZE, and TX.

```r
fit.extra.1 <- coxph(Surv(INTTIME, EVENT)~NUM+SIZE+TX+NUM*TX, data=bladder)
fit.extra.1
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ NUM + SIZE + TX + NUM *
##     TX, data = bladder)
##
##           coef exp(coef) se(coef)     z     p
## NUM     0.0533    1.0548   0.0830  0.64 0.520
## SIZE   -0.0374    0.9633   0.0720 -0.52 0.603
## TX     -0.7004    0.4964   0.3530 -1.98 0.047
## NUM:TX  0.1404    1.1507   0.1051  1.34 0.182
##
## Likelihood ratio test=10.28  on 4 df, p=0.04
## n= 191, number of events= 112
```

```r
fit.extra.2 <- coxph(Surv(INTTIME, EVENT)~NUM+SIZE+TX, data=bladder)
fit.extra.2
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ NUM + SIZE + TX, data = bladder)
##
##         coef exp(coef) se(coef)     z      p
## NUM   0.1362    1.1459   0.0504  2.70 0.0069
## SIZE -0.0144    0.9857   0.0695 -0.21 0.8357
## TX   -0.3185    0.7273   0.2031 -1.57 0.1169
##
## Likelihood ratio test=8.44  on 3 df, p=0.04
## n= 191, number of events= 112
```

We then perform the likelihood ratio test based on the results above:

```
lrt.num.extra = 2*(fit.extra.1$loglik[2]-fit.extra.2$loglik[2])
lrt.num.extra
```

```
## [1] 1.842129
```

```
pchisq(lrt.num.extra,df=1,lower.tail=FALSE)
```

```
## [1] 0.1747016
```

The p-value is $0.1747016 > \alpha = 0.05$, and so we accept the null hypothesis and conclude that NUM*TX is not a significant covariate.

Although we are not yet applying the assumption of recurrence, by nature of the recurrent data, because we cannot predict what affects recurrence of bladder cancer for a patient who does experiences recurrence, we would want to stratify the one covariate which has been tested to have significantly affect time, which is the NUM covariate. We perform stratification below, starting with the non-interaction term and then proceeding with the interaction:

```
fit.extra.strata.noint <- coxph(Surv(INTTIME, EVENT)~strata(NUM)+TX, data=bladder)
summary(fit.extra.strata.noint)
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ strata(NUM) + TX, data = bladder)
##
##   n= 191, number of events= 112
##
##       coef exp(coef) se(coef)      z Pr(>|z|)
## TX -0.3843    0.6809   0.2209 -1.739    0.082 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##    exp(coef) exp(-coef) lower .95 upper .95
## TX    0.6809      1.469    0.4416      1.05
##
## Concordance= 0.536  (se = 0.046 )
## Rsquare= 0.016   (max possible= 0.971 )
## Likelihood ratio test= 3.16  on 1 df,   p=0.08
## Wald test            = 3.03  on 1 df,   p=0.08
## Score (logrank) test = 3.06  on 1 df,   p=0.08
```

```
fit.extra.strata.int <- coxph(Surv(INTTIME, EVENT)~strata(NUM)*TX, data=bladder)
summary(fit.extra.strata.int)
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ strata(NUM) * TX, data = bladder)
```

```
##
##   n= 191, number of events= 112
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## TX                    -0.4809    0.6182   0.3197 -1.504    0.133
## strata(NUM)NUM=2:TX -0.1276    0.8802   0.7127 -0.179    0.858
## strata(NUM)NUM=3:TX -0.1475    0.8629   0.6736 -0.219    0.827
## strata(NUM)NUM=4:TX  0.4414    1.5549   0.8223  0.537    0.591
## strata(NUM)NUM=5:TX  0.8205    2.2716   0.7550  1.087    0.277
## strata(NUM)NUM=6:TX      NA        NA   0.0000     NA       NA
## strata(NUM)NUM=8:TX  0.3521    1.4220   1.2704  0.277    0.782
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## TX                      0.6182     1.6175    0.3304     1.157
## strata(NUM)NUM=2:TX     0.8802     1.1362    0.2177     3.558
## strata(NUM)NUM=3:TX     0.8629     1.1589    0.2304     3.231
## strata(NUM)NUM=4:TX     1.5549     0.6431    0.3103     7.792
## strata(NUM)NUM=5:TX     2.2716     0.4402    0.5172     9.978
## strata(NUM)NUM=6:TX         NA         NA        NA        NA
## strata(NUM)NUM=8:TX     1.4220     0.7032    0.1179    17.150
##
## Concordance= 0.541  (se = 0.046 )
## Rsquare= 0.025   (max possible= 0.971 )
## Likelihood ratio test= 4.93  on 6 df,   p=0.6
## Wald test            = 4.56  on 6 df,   p=0.6
## Score (logrank) test = 4.66  on 6 df,   p=0.6
```

We then try to see if the interaction term (with stratification) is significant or not using likelihood ratio test.

```
lrt.extra.int = 2*(fit.extra.strata.int$loglik[2]-fit.extra.strata.noint$loglik[2])
pchisq(lrt.extra.int,df=6,lower.tail = FALSE)
```

```
## [1] 0.9396426
```

The p-value is large so therefore the interaction term is not significant. So therefore, we can for now base our model on our previous analysis, which did not include the interaction term.

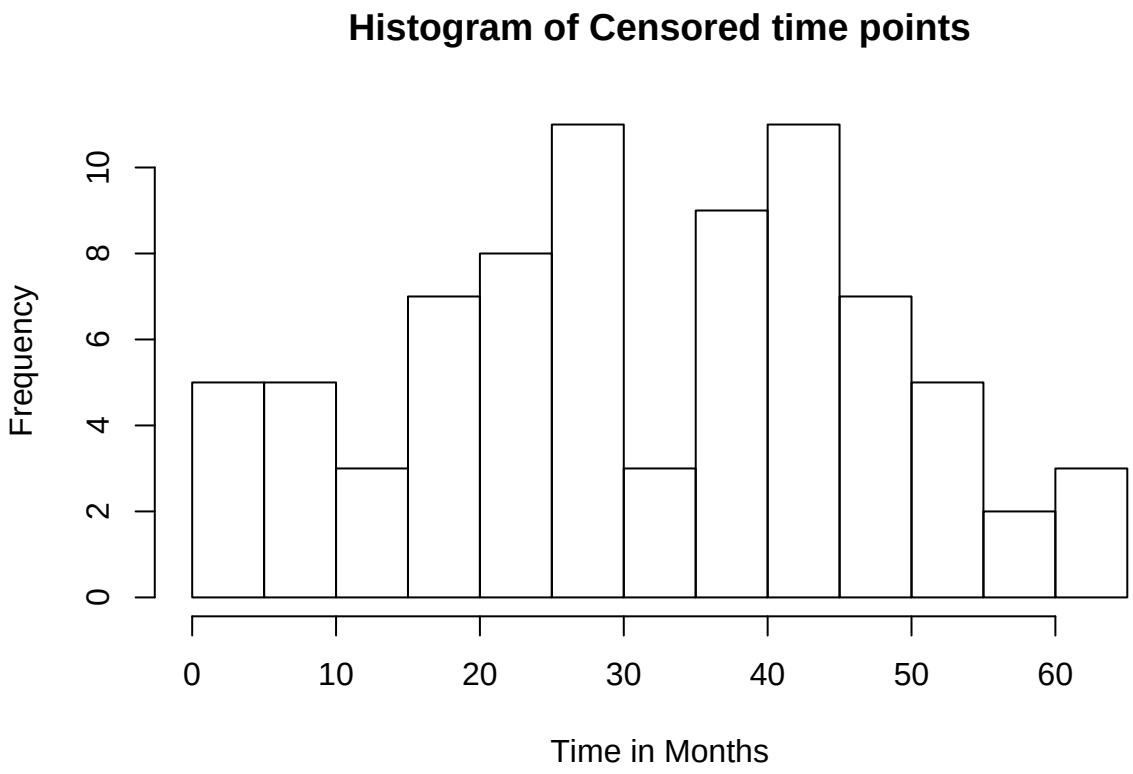
Part Two: Recurrence

We now apply how the bladder data has recurrent data. We observe that for every patient in the data set, after they are finished receiving treatments or there no longer exists records of the patients, the patient's data gets censored, as indicated by the EVENT variable being equal to 0. Because of recurrence, our survivability distribution may look differently from the survivability distribution of our data when we had assumed no recurrence. Because of recurrence, we therefore have to take into consideration the START and STOP variables, START indicating at what time the patient starts his treatment, and STOP indicating at what time the patient ends his treatment. Later when we make use of the survfit function, making use of the START and STOP variables helps us to take into consideration recurrence in our code, especially when we make use of the Surv() function in R.

We first take a look at the frequency distribution of the times at which the patients are censored.

```
ordered.censored.times <- bladder$STOP[bladder$EVENT == 0]
ordered.censored.times <- sort(ordered.censored.times)

hist(ordered.censored.times, breaks = 14, main = "Histogram of Censored time points", xlab = "Time in M
```

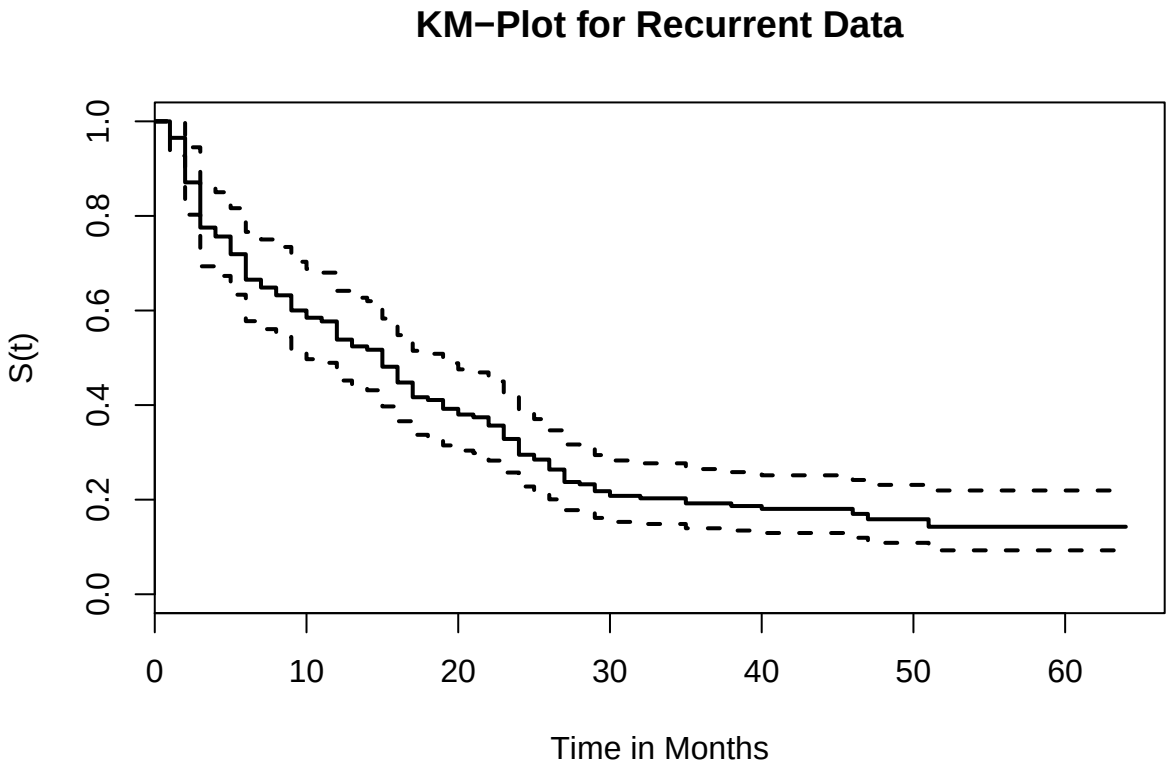


Above, we see that most of the patients are being censored during the first 40 months of the observation period, and this matches the results that we've obtained from the original KM-plot, in which much of the decrease in the survivability times of bladder cancer patients had occurred in the first 40 months.

We now want to plot the KM-plot while taking into account recurrence. However, we notice that the patient with ID = 1 has a START and STOP time of 0. Because we want to take into account recurrence of bladder cancer, we perform a model setup by making slight revisions to the dataset by changing the STOP time of the patient to be 1.

```
bladder$STOP[1] <- 1

plot(survfit(Surv(START,STOP,EVENT)~1,data=bladder),lwd=2,xlab="Time in Months",ylab="S(t)",
     main = "KM-Plot for Recurrent Data")
```



From the KM-plot above, we see that there is still a convergence to a survivability probability as there was in the original KM-plot. However, in the previous KM-plot, convergence appeared to happen sooner starting from time = 40 whereas in the above recurrent plot, convergence seems to happen at time = 50. We also observe that the probability at which the recurrent KM-plot converges at is 0.2 as opposed to 0.3 in the original KM-plot. However, it still appears that within the first year, half of the patients are censored or dropped from the experiment.

We now perform quantile estimation to see at which times the KM-plot reaches a certain point of survivability.

```
quantile(survfit(Surv(START,STOP,EVENT)~1,data=bladder), probs = c(.1, .3, .5), conf.int = F)
```

```
## 10 30 50
##  2  6 15
```

From the above results, we can see that at time 2, about 90% of the patients remain in the experiment. At time 6, about 70% of the patients remain in the experiment. And finally, at time 15, about 50% of the patients remain in the experiment.

We observe that in general, when taking recurrence into account, the patients appear to have longer survivability times than that of the patients of when we assume that they are all independent.

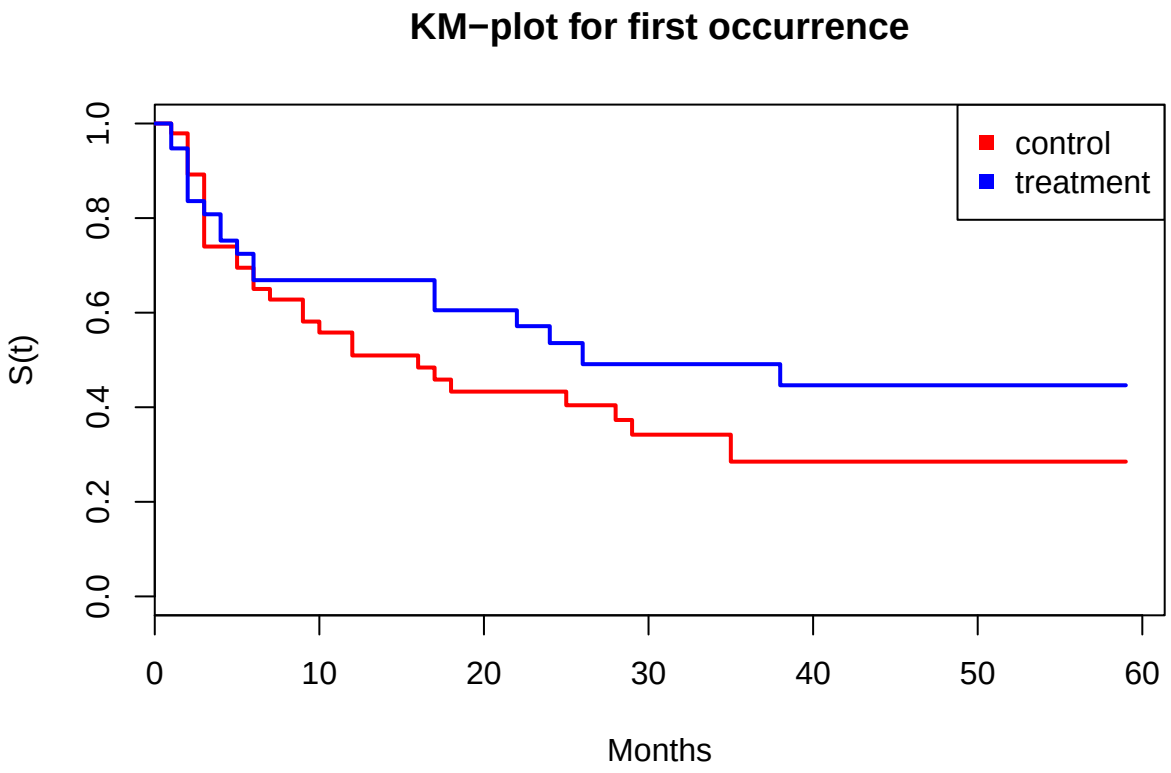
```
max(bladder$INTERVAL)
```

```
## [1] 5
```

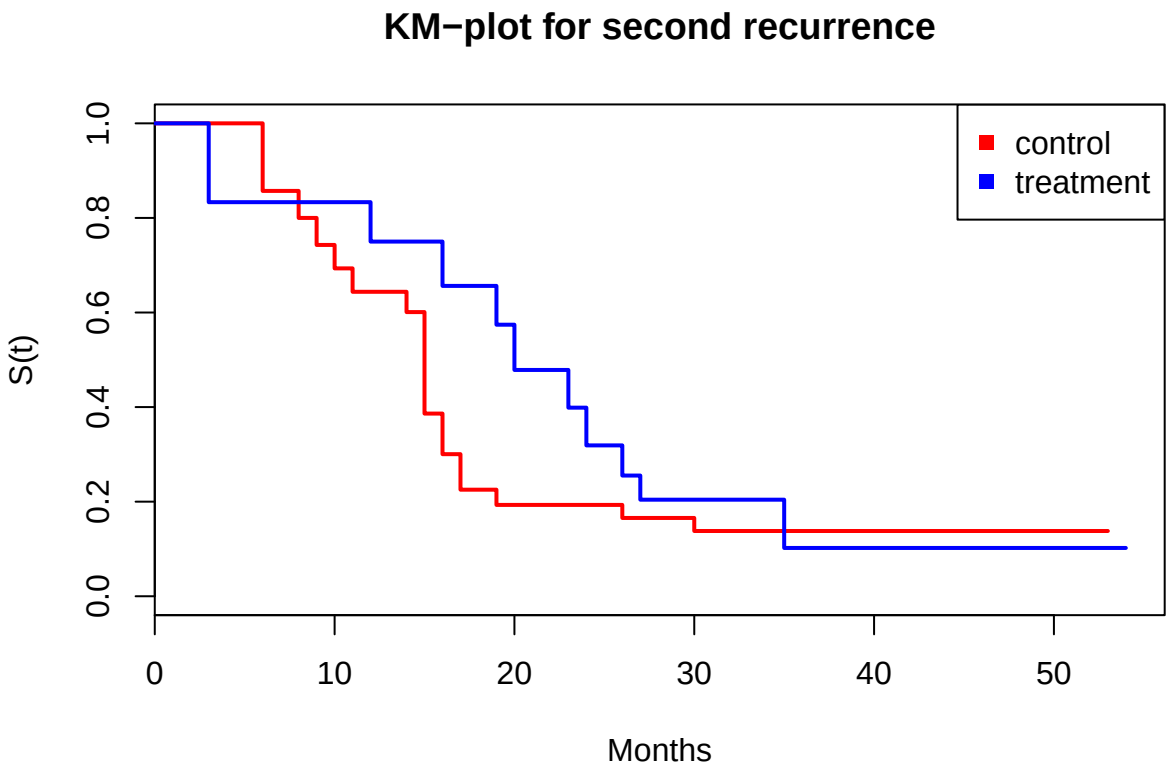
From above, we can see that the maximum number of recurrences for bladder cancer is 5. We then proceed to plot KM-plots for different number of recurrences for the patients given that we have two types of treatment groups.

```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
```

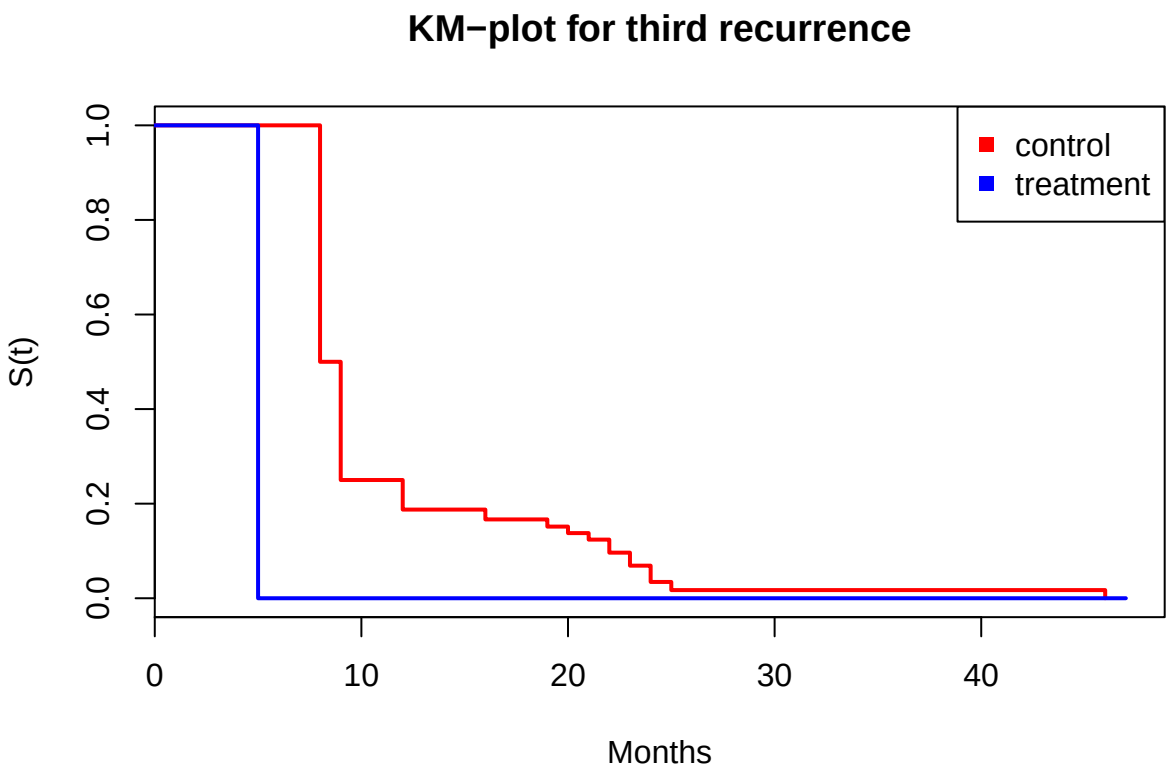
```
              subset=(bladder$INTERVAL == "1")),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for first occurrence" )
legend("topright", legend = c("control", "treatment"), pch = 15, col = c(2,4))
```



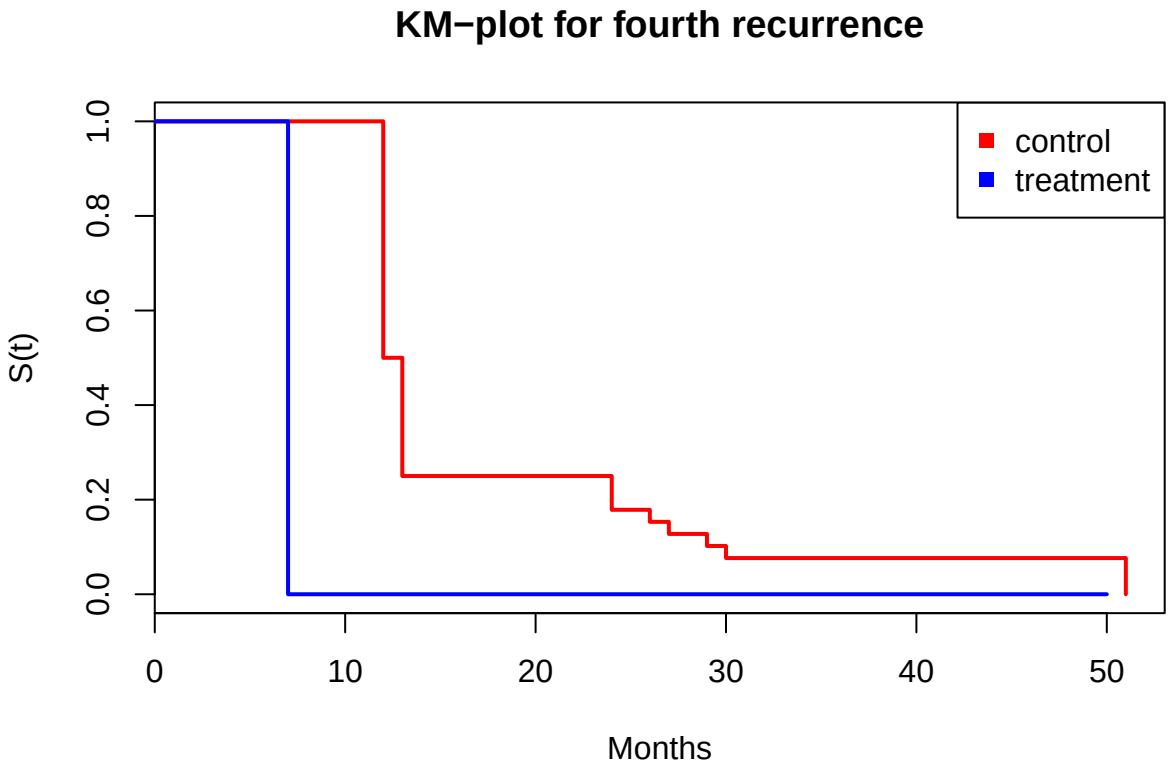

```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
             subset=(bladder$INTERVAL == "2")),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for second recurrence" )
legend("topright", legend = c("control", "treatment"), pch = 15, col = c(2,4))
```



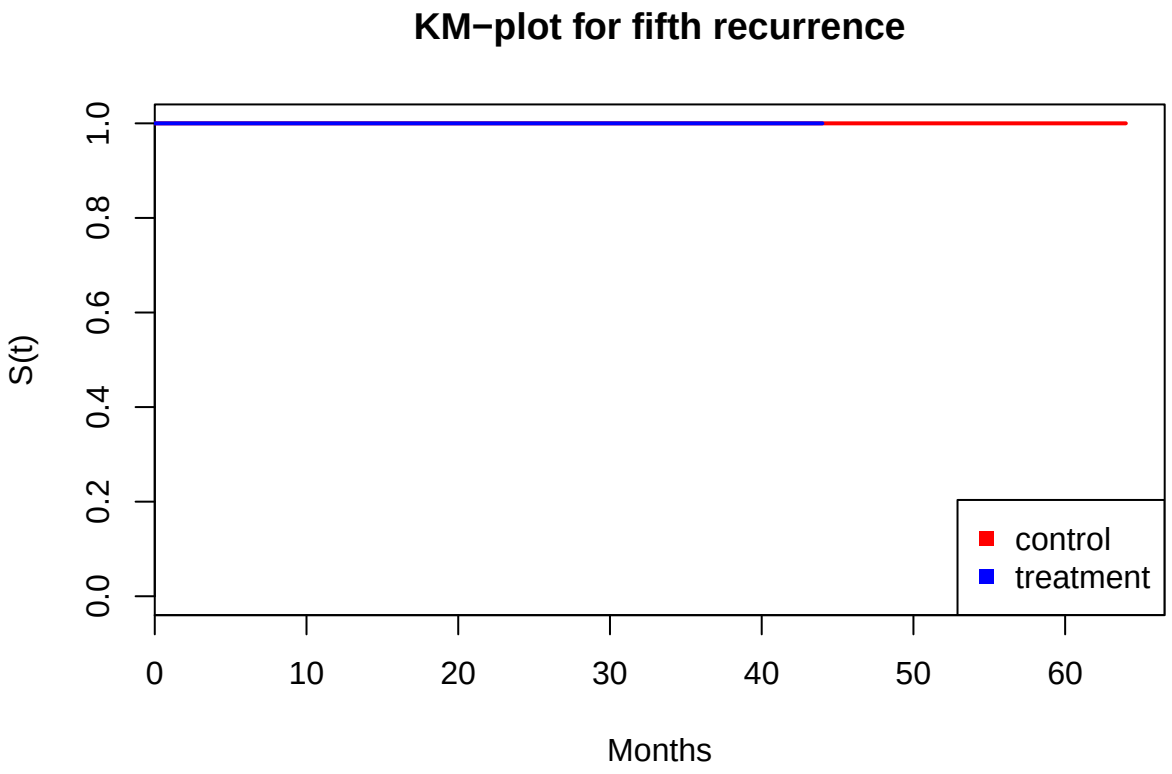

```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
             subset=(bladder$INTERVAL == "3")),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for third recurrence" )
legend("topright", legend = c("control", "treatment"), pch = 15, col = c(2,4))
```




```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
             subset=(bladder$INTERVAL == "4")),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for fourth recurrence" )
legend("topright", legend = c("control", "treatment"), pch = 15, col = c(2,4))
```




```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
             subset=(bladder$INTERVAL == "5")),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for fifth recurrence" )
legend("bottomright", legend = c("control", "treatment"), pch = 15, col = c(2,4))
```

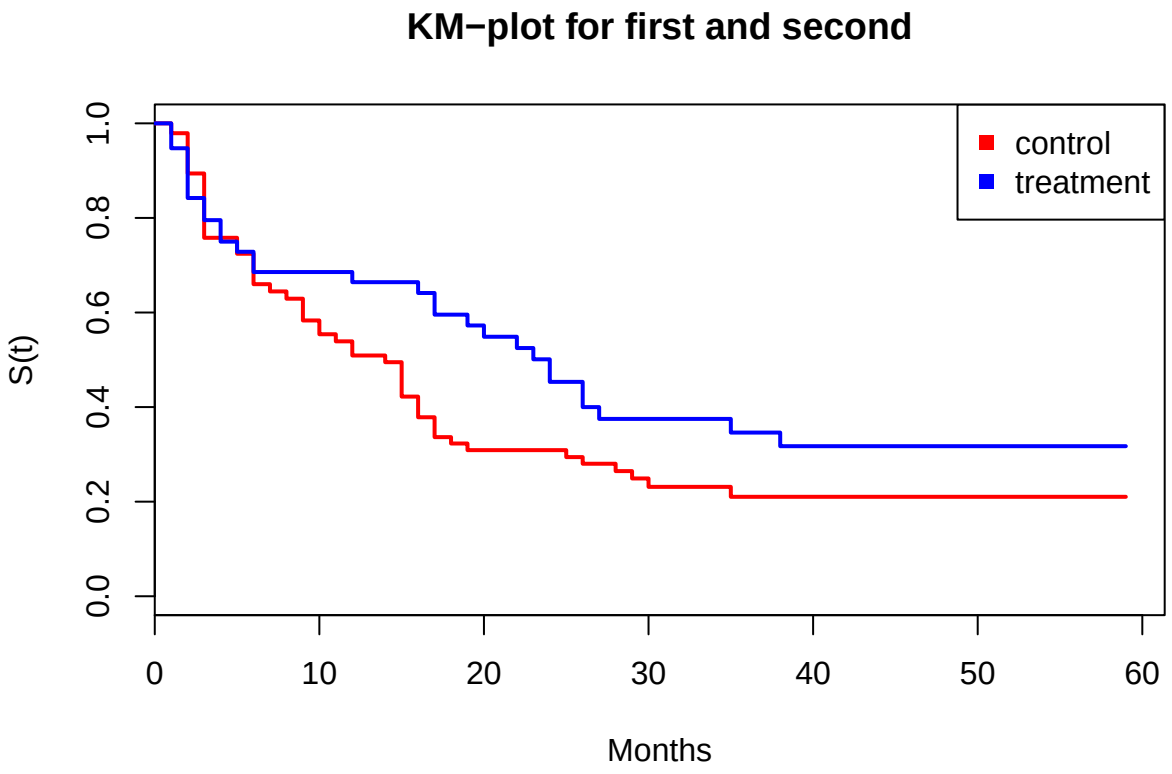


For the KM-plot of the 5th recurrence of bladder cancer of patients, the reason the survivability curves for both of the treatment groups is equal to 1 is because at that point, the patients' data is censored. In other words, the patients whose data is censored leaves the experiment; however, by nature of censorship, we assume that the patients have survived because we would still want to make use of the censored data for our survivability distribution. Therefore, the survivability of these patients is going to be constantly 1.

From the above plots, we can see that the survivability curves for 1 occurrence and 2nd recurrence are similar to each other. There is not enough data for the 3rd, 4th, and 5th recurrences, so therefore, we would want to group the data together: have the 1st and 2nd recurrence data be in one group, and the 3rd, 4th, and 5th recurrence data be in another group.

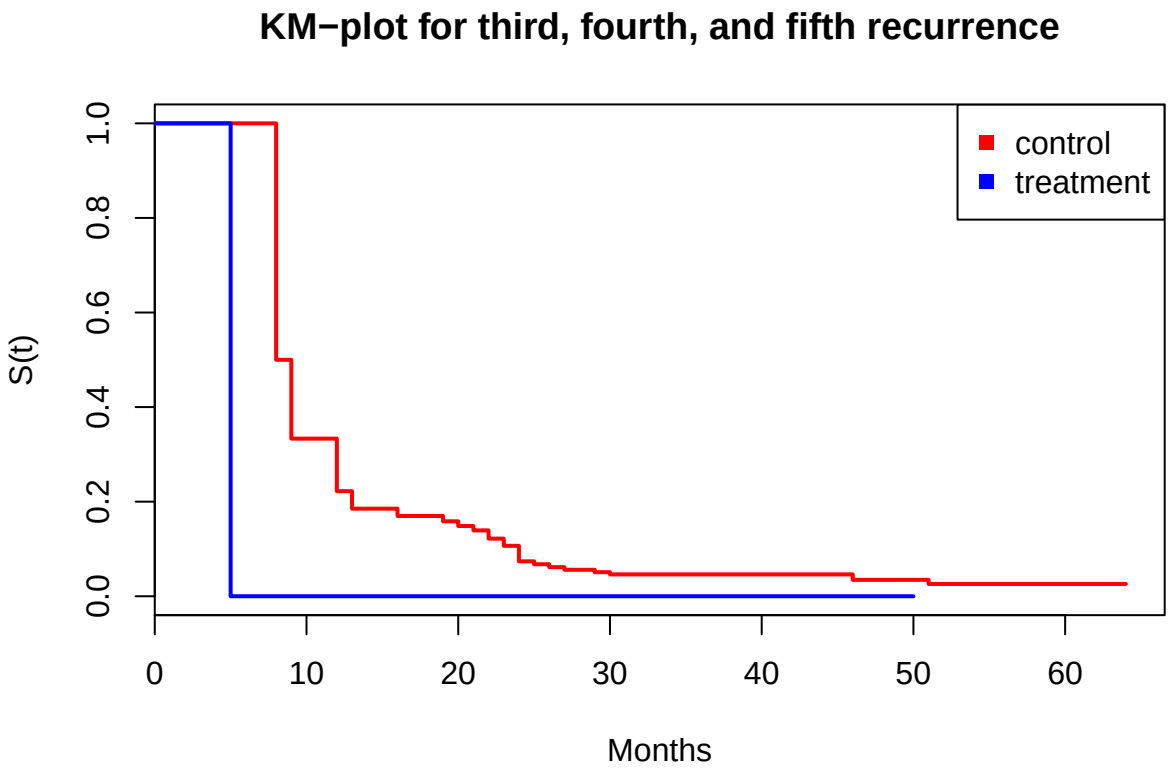
```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
             subset=(bladder$INTERVAL < 3)),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for first and second" )

legend("topright", legend = c("control", "treatment"), pch = 15, col = c(2, 4))
```




```
plot(survfit(Surv(START,STOP,EVENT)~TX,
             data=bladder,
             subset=(bladder$INTERVAL > 2)),
      lwd=2, col=c(2,4),xlab="Months",ylab="S(t)",
     main = "KM-plot for third, fourth, and fifth recurrence" )

legend("topright", legend = c("control", "treatment"), pch = 15, col = c(2,4))
```



From the above plots: plot of group 1, which includes data on the first and second recurrence, and group 2, which includes data on the third, fourth, and fifth recurrence, we observe there is significant difference between the two KM-plots. Therefore, we'd want to explore further as to see whether or not recurrence has a significant effect on time. We make use of the INTERVAL variable to determine those significant effects.

COX PH using the INTERVAL variable

We are now curious about fitting a COX PH model for the recurrence of the patients in addition to continuing our analysis of the two treatment groups. Because we are curious about how recurrence in data affects the survival analysis that we perform, we would want to involve the INTERVAL covariate. We first fit a Cox PH model which includes both the TX and INTERVAL covariates. Again in our Surv function, we make sure to make use of the START and STOP variables in order to take into consideration recurrence of bladder cancer.

```
fit.interval <- coxph(Surv(START, STOP, EVENT) ~ TX + INTERVAL, data = bladder)
anova(fit.interval)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(START, STOP, EVENT)
## Terms added sequentially (first to last)
##
##           loglik  Chisq Df Pr(>|Chi|)
## NULL     -464.34
## TX       -462.99 2.7027  1   0.100180
## INTERVAL -458.09 9.8047  1   0.001741 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above, we can see that INTERVAL has a p-value of $0.001741 < \alpha = 0.05$, which indicates that it is a significant covariate for the model. We also include the TX covariate because it is sufficiently small. We oftentimes cannot predict how recurrence of bladder cancer affects the next recurrence of the cancer, and

so we want to see if the INTERVAL variable has an effect on the survivability time of the patients via the likelihood ratio test. We therefore fit a Cox PH reduced model containing only the TX covariate.

```
fit.interval.reduced <- coxph(Surv(START, STOP, EVENT) ~ TX, data = bladder)
anova(fit.interval.reduced)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(START, STOP, EVENT)
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -464.34
## TX   -462.99 2.7027  1     0.1002
```

```
lrt.interval = 2*(fit.interval$loglik[2]-fit.interval.reduced$loglik[2])
pchisq(lrt.interval,df=1,lower.tail = FALSE)
```

```
## [1] 0.001740653
```

The p-value is $0.001740653 < \alpha = 0.05$ and so therefore, we reject the null hypothesis and conclude that INTERVAL has a significant impact on the survivability time, which means that recurrence does have a significant impact on time. This also means that our previous model of fit1 is incomplete, for we do not take into consideration of the recurrence of bladder cancer. We therefore proceed by fitting a model which takes into account recurrence by testing out the main covariates, NUM, SIZE, and TX.

Cox PH Regression for Recurrence

We again perform backwards elimination for covariates NUM, SIZE, and TX to see which model fix best to explain which covariates have significant effects on time. We will have our full model (which includes covariates NUM, SIZE, and TX). We perform analysis on variance below on the cox PH model:

```
fit1.2 <- coxph(Surv(START, STOP, EVENT)~NUM+SIZE+TX, data=bladder)
anova(fit1.2)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(START, STOP, EVENT)
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -464.34
## NUM  -459.36 9.9622  1   0.001598 **
## SIZE -459.20 0.3334  1   0.563692
## TX   -457.01 4.3697  1   0.036584 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results above, we can see that we have the following variables in order of significance based on the p-values: NUM, TX, and SIZE. We fit two more models, fit2.2 (which contains NUM and TX covariates), and fit3.2 (which contains NUM covariate), and perform AIC to see which model is the best fit for our Cox Regression model.

```
fit2.2 <- coxph(Surv(START, STOP, EVENT)~NUM+TX, data=bladder)

fit3.2 <- coxph(Surv(START, STOP, EVENT)~NUM, data=bladder)

AIC(fit1.2)
```

```
## [1] 920.0206
```

```
AIC(fit2.2)
```

```
## [1] 918.3689
```

```
AIC(fit3.2)
```

```
## [1] 920.7236
```

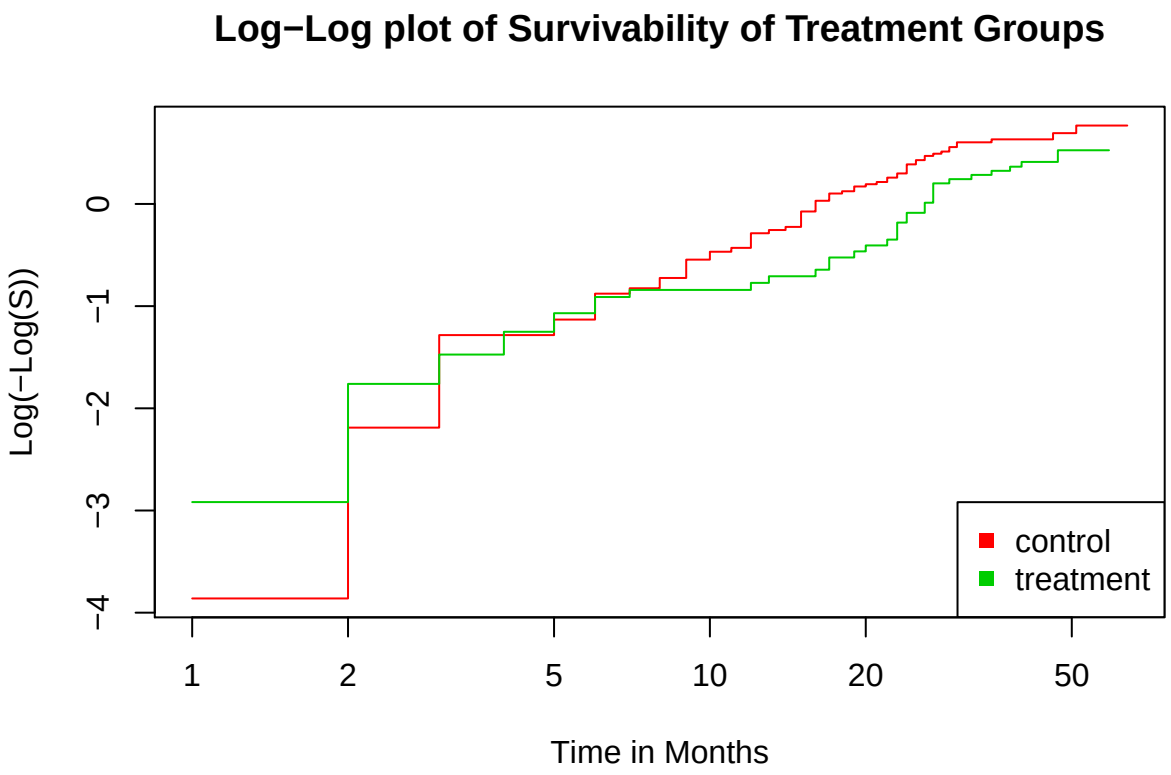
We see from the above results that fit2.2 is the best model, which contains both the NUM and TX covariates. Regardless of whether we take into consideration recurrence, it seems that our choice of covariates still holds.

Comparison: Treatment groups under Recurrence

We now look at the two treatment groups under the recurrence assumption. We first look at the log-log plot of the treatment groups to see if the Cox PH assumption holds for modelling the effects of the treatment groups.

```
split.fit2 <- survfit(Surv(START, STOP, EVENT)~TX, data=bladder)
plot(split.fit2, fun="cloglog", col=c(2,3), xlab="Time in Months", ylab="Log(-Log(S))",
     main = "Log-Log plot of Survivability of Treatment Groups")

legend("bottomright", legend = c("control", "treatment"), pch = 15, col = 2:3)
```



From the above plot, because there are intersections between the two plots of the control and treatment groups, the coxPH assumption does not hold for modelling the treatment group.

However, we decide to test the Cox PH assumption on the treatment covariate and see if our observations from the log-log plot contradict the results from applying the Cox PH assumption to the treatment groups under recurrence.

```
blad2 = coxph(Surv(START, STOP, EVENT)~TX, data = bladder)
blad2
```

```
## Call:
## coxph(formula = Surv(START, STOP, EVENT) ~ TX, data = bladder)
##
##      coef exp(coef) se(coef)     z   p
## TX -0.320     0.726    0.197 -1.62 0.1
##
## Likelihood ratio test=2.7  on 1 df, p=0.1
## n= 191, number of events= 112
```

From the above coxph function, we can estimate the hazard proportion between the control and treatment group: $(h_0(t)e^{\beta x})/h_0(t) = e^{\beta X_1}$. From the above result, we obtain $e^{\beta} = 0.726$, which is the value of our proportion. For one unit increase in X_1, the hazard increases by a factor of e^{β}, for $e^{\beta} < 1$. The risk of recurrence decreases by 17.4% with each unit increase in X_1 (the beta corresponding to TX is negative, so the hazard ratio decreases with time).

```
exp(confint(blad2, level = 0.95))
```

```
##        2.5 %   97.5 %
## TX 0.4929658 1.068939
```

The 95% confidence interval for the hazards ratio for treatment is (0.4929658, 1.068939).

We now use the cox.zph function to perform a test to see if the model, fit2.2, is significantly divergent from the proportional hazards model

```
cox.zph(fit2.2,global = FALSE) #NUM
```

```
##        rho chisq     p
## NUM 0.0430 0.166 0.684
## TX  0.0345 0.126 0.723
```

Our p-values are greater than $\alpha = 0.05$, so there is insufficient evidence to suggest that the PH assumption does not hold. This is consistent with our results from the graphs and previous tests. We conclude that the Cox PH is reasonable because there is not significant evidence for us to reject that assumption.

Another type of Recurrent Event Analysis: GAP Time Model

Because we are working with a data set which takes into account recurrence, we would want to make use of the GAP Time (GT) Model. When using the GT model, we would want to make use of the variable INTTIME, for there may be interest in the time interval, which is the time from the previous event to the next recurrent event. With GT, when the first event occurs, there isn't a change in the risk set for recurrent events.

Below, we have our GT model:

```
coxph(Surv(INTTIME,EVENT)~ INTERVAL + TX + cluster(ID), data=bladder)
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ INTERVAL + TX + cluster(ID),
##     data = bladder)
##
##              coef exp(coef) se(coef) robust se     z    p
## INTERVAL  0.0185    1.0187   0.0734    0.0569  0.33 0.74
```

```
## TX       -0.2067    0.8133   0.1997    0.2156 -0.96 0.34
##
## Likelihood ratio test=1.24  on 2 df, p=0.5
## n= 191, number of events= 112
```

From above, $\beta_{TX} = -0.2067$ and the 95% confidence interval for β_{TX}, using nonrobust standard errors, is: $[-0.598112, 0.184712]$. Our point estimate of the hazard ratio for TX is: 0.81326. Our 95% confidence interval for the hazard ratio is: $[0.54985, 1.20287]$.

We also take into consideration what the hazard ratio for TX would be when we decide to stratify the INTERVAL variable.

```
coxph(Surv(INTTIME,EVENT)~ strata(INTERVAL) + TX + cluster(ID), data=bladder)
```

```
## Call:
## coxph(formula = Surv(INTTIME, EVENT) ~ strata(INTERVAL) + TX +
##     cluster(ID), data = bladder)
##
##      coef exp(coef) se(coef) robust se     z    p
## TX -0.163     0.849    0.202     0.219 -0.75 0.46
##
## Likelihood ratio test=0.66  on 1 df, p=0.4
## n= 191, number of events= 112
```

From above, $\beta_{TX} = -0.163$, and the 95% confidence interval for β_{TX}, using nonrobust standard errors, is: $[-0.55892, 0.23292]$. Our point estimate of the hazard ratio for TX is: 0.84959. Our 95% confidence interval for the hazard ratio is: $[0.57183, 1.26228]$.

FINAL MODEL

We conclude our analysis of the bladder cancer dataset with our final model. Inititally, when we were deriving our model, we assumed that there didn't exist recurrence of bladder cancer in patients, and have determined that NUM and TX were to be used for our main covarariates.

However, we later found that recurrence has an effect on the survivability time, and so therefore, while taking into account recurrence, we decided to fit a new model, which still shows that NUM and TX are the significant covariates which have impacts on the time. We also found from the coxzph function that this model satisfies the Cox PH Assumption, and so NUM and TX are valid covariates for our Cox PH model for recurrence. Below, we have information about the covariates from our final model:

```
fit2.2
```

```
## Call:
## coxph(formula = Surv(START, STOP, EVENT) ~ NUM + TX, data = bladder)
##
##        coef exp(coef) se(coef)     z       p
## NUM  0.1701    1.1854   0.0465  3.66 0.00025
## TX  -0.4113    0.6628   0.2003 -2.05 0.04002
##
## Likelihood ratio test=14.32  on 2 df, p=8e-04
## n= 191, number of events= 112
```

From our results, we have determined that $\beta_{\text{NUM}} = 0.1701$, and that the 95% confidence interval for β_{NUM} is: $0.1701 \pm 1.96(0.0465)$, which is: $[0.07896, 0.26124]$. Our hazard ratio is then: $e^{\beta_{\text{NUM}}} \approx e^{0.1701} = 1.18542$, and our 95% confidence interval for the hazard ratio is: $[1.08216, 1.29854]$.

We have also determined that $\beta_{\text{TX}} = -0.4113$, and that the 95% confidence interval for β_{TX} is: $[-0.803888, -0.018712]$. Our hazard ratio is then: $e^{\beta_{\text{TX}}} \approx 0.66279$, and our 95% confidence interval for the hazard ratio is: $[0.44759, 0.98146]$.

We see from our results that $e^{\beta_{\text{NUM}}} > 1$, which means that there is an increase in the hazard for every unit increase in X_{NUM}. Therefore, the length of survival time decreases because of the increase in hazard. We also see that $e^{\beta_{\text{TX}}} < 1$, which means that there is a decrease in the hazard for every unit increase in X_{TX}. Therefore, the length of survival time increases because of the decrease in the hazard.

Acknowledgements

We would like to thank Professor Andrew Carter and Nhan Huynh for giving us guidance for how to proceed with the project.

Reference

Byar, D. 1980. The Veterans Administration study of chemopro- phylaxis for recurrent stage I bladder tumors: Comparisons of placebo, pyridoxine, and topical thiotepa. In Bladder Tumors and Other Topics in Urological Oncology. New York: Plenum, 363–370.