Introduction

The “Colon” dataset, provided by the Survival package in R, derives from the trials of adjuvant chemotherapy for colon cancer. The patients from the trial were randomly assigned to no further treatment or to an adjuvant treatment with either levamisole or levamisole plus fluorouracil (5-FU). The dataset contains two records per patient, one for recurrence and the other for death, resulting in a total of 1,858 observations. The attributes within the study are as follows:

1. Rx: Treatment - Obs(ervation), Lev(amisole), Lev(amisole)+5-FU
2. Sex: male and female
3. Age: in years
4. Obstruct: obstruction of colon by tumor
5. Perfor: perforation of colon
6. Adhere: adherence to nearby organs
7. Nodes: number of lymph nodes with detectable cancer
8. Time: days until event or censoring
9. Status: censoring status
10. Differ: differentiation of tumor (1=well, 2=moderate, 3=poor)
11. Extent: extent of local spread (1=submucosa, 2=muscle, 3=serosa, 4=contiguous structures)
12. Surg: time from surgery to registration (0=short, 1=long)
13. Node4: more than 4 positive lymph nodes
14. Etype: event type: 1=recurrence, 2=death

The purpose of our project is to establish appropriate cox proportional hazards models for the treatments and to use the developed models to answer our research questions. Our project focused on examining which factors had an effect on time to recurrence and time to mortality as well as whether the treatments, levamisole or levamisole plus fluorouracil (5-FU), were effective throughout the course of the study.

Questions of Interest

For this research paper, we seek to answer the following questions:

> **Question 1:** Do the treatments, Levamisole and Levamisole+5-FU, help improve the survival rate in colon cancer patients?
>
> **Question 2:** Do the treatments, Levamisole and Levamisole+5-FU, delay the recurrence of colon cancer?
>
> **Question 3:** Do the treatments, Levamisole and Levamisole+5-FU, help improve the survival rate after the recurrence of colon cancer?

Survival Analysis for the Event of Mortality

Model Building Process

Examining the data

Recall that there are two records for each patient indicated by the event type (etype) variable, where etype == 1 refers to the event of a recurrence and etype == 2 indicates death. In order to answer our first research question, which is to study the time until death, we must create a marginal model by subsetting the colon data to only include the event of mortality. To get an overview of the mortality subset we use the survfit function and plot the Kaplan-Meier Estimate between the three different treatments.

```
#subset death data
colon.death <-subset(colon, etype ==2)

death.fit <-survfit(Surv(time,status) ~rx, data =colon.death)

ggsurvplot(death.fit, conf.int =F,
        title ="Kaplan-Meier Curve for Colon Cancer Mortality \nby Treatment",
        xlab ="Time (until death) \nin Days")
```

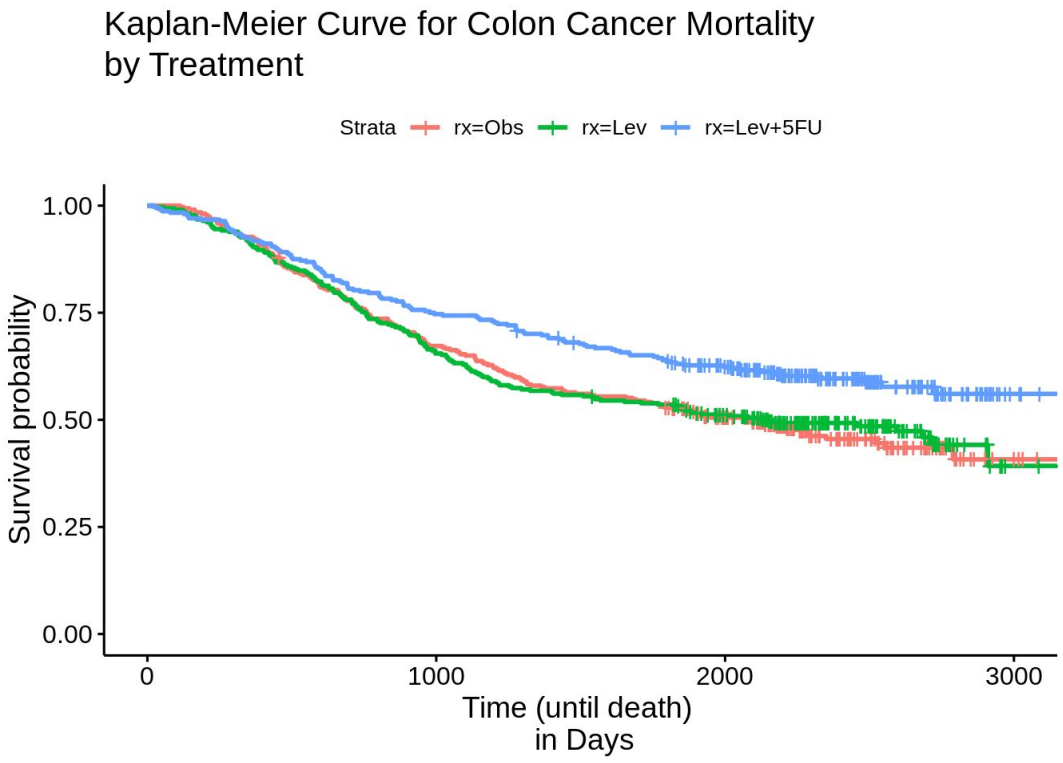

Figure 1: Survival probability for colon cancer patients based on treatment.

Based on Figure 1, there is some indication that patients who received the adjuvant treatment with levamisole plus fluorouracil (Lev+5Fu) have a higher survival probability than patients with no further treatment and patients who received the treatment with levamisole alone.

Moreover, it is important to note that there are columns that contain NA values. Out of 929 observations, 41 of them contain NA values in at least one column. Since observations that contain NA values make up only 4% of our data, we decided that removing them wouldn't cause a big effect on the variable selection process. By removing observations with NA values, created a new mortality dataset colon.death1 which contains 888 observations.

```
#removing NA values
colon.death1 <-na.omit(colon.death)
```

Variable Selection

We now use forward selection with Bayesian Information Criterion (BIC) to determine the covariates that best represent an appropriate cox proportional hazards model for the event of death. Within each step, we chose the model that has the lowest BIC value.

```
d.model1 <-coxph(Surv(time, status) ~sex, data=colon.death1)
d.model2 <- coxph(Surv(time, status) ~ age, data=colon.death1)
d.model3 <-coxph(Surv(time, status) ~obstruct, data=colon.death1)
d.model4 <-coxph(Surv(time, status) ~perfor, data=colon.death1)
d.model5 <-coxph(Surv(time, status) ~adhere, data=colon.death1)
d.model6 <-coxph(Surv(time, status) ~nodes, data=colon.death1)
d.model7 <-coxph(Surv(time, status) ~differ, data=colon.death1)
d.model8 <-coxph(Surv(time, status) ~extent, data=colon.death1)
d.model9 <-coxph(Surv(time, status) ~surg, data=colon.death1)
d.model10 <-coxph(Surv(time, status) ~node4, data=colon.death1)

BIC(d.model1, d.model2, d.model3, d.model4, d.model5, d.model6, d.model7, d.model8, d.model9,
d.model10)
##          df     BIC
## d.model1  1 5541.880
## d.model2  1 5541.408
## d.model3  1 5537.618
## d.model4  1 5541.525
## d.model5  1 5536.616
## d.model6  1 5471.321
## d.model7  2 5535.900
## d.model8  3 5526.573
## d.model9  1 5538.031
## d.model10 1 5458.201
```

The model with the smallest BIC value in this step is d.model10 which contains covariate node4.

```
d.model10.1<-coxph(Surv(time, status) ~node4 +sex, data=colon.death1)
d.model10.2<-coxph(Surv(time, status) ~node4 +age, data=colon.death1)
d.model10.3<-coxph(Surv(time, status) ~node4 +obstruct, data=colon.death1)
```

```
d.model10.4<-coxph(Surv(time, status) ~node4 +perfor, data=colon.death1)
d.model10.5<-coxph(Surv(time, status) ~node4 +adhere, data=colon.death1)
d.model10.6<-coxph(Surv(time, status) ~node4 +nodes, data=colon.death1)
d.model10.7<-coxph(Surv(time, status) ~node4 +differ, data=colon.death1)
d.model10.8<-coxph(Surv(time, status) ~node4 +extent, data=colon.death1)
d.model10.9<-coxph(Surv(time, status) ~node4 +surg, data=colon.death1)

BIC(d.model10.1, d.model10.2, d.model10.3, d.model10.4, d.model10.5, d.model10.6, d.model10.7,
d.model10.8, d.model10.9)
##              df      BIC
## d.model10.1 2 5464.237
## d.model10.2 2 5461.289
## d.model10.3 2 5458.564
## d.model10.4 2 5463.690
## d.model10.5 2 5459.417
## d.model10.6 2 5455.527
## d.model10.7 3 5463.982
## d.model10.8 4 5455.598
## d.model10.9 2 5458.843
```

The model with the smallest BIC value in this step is d.model10.6 which contains covariate node4 and nodes.

```
d.model10.6.1<-coxph(Surv(time, status) ~node4 +nodes +sex, data=colon.death1)
d.model10.6.2<-coxph(Surv(time, status) ~node4 +nodes +age, data=colon.death1)
d.model10.6.3<-coxph(Surv(time, status) ~node4 +nodes +obstruct, data=colon.death1)
d.model10.6.4<-coxph(Surv(time, status) ~node4 +nodes +perfor, data=colon.death1)
d.model10.6.5<-coxph(Surv(time, status) ~node4 +nodes +adhere, data=colon.death1)
d.model10.6.6<-coxph(Surv(time, status) ~node4 +nodes +differ, data=colon.death1)
d.model10.6.7<-coxph(Surv(time, status) ~node4 +nodes +extent, data=colon.death1)
d.model10.6.8<-coxph(Surv(time, status) ~node4 +nodes +surg, data=colon.death1)

BIC(d.model10.6.1, d.model10.6.2, d.model10.6.3, d.model10.6.4, d.model10.6.5, d.model10.6.6,
d.model10.6.7, d.model10.6.8)
##              df      BIC
## d.model10.6.1  3 5461.590
## d.model10.6.2  3 5458.576
## d.model10.6.3  3 5455.818
## d.model10.6.4  3 5461.066
## d.model10.6.5  3 5456.915
## d.model10.6.6  4 5462.327
## d.model10.6.7  5 5454.430
## d.model10.6.8  3 5455.309
```

The model with the smallest BIC value in this step is d.model10.6.7 which contains covariate node4, nodes and extent.

```
d.model10.6.7.1<-coxph(Surv(time, status) ~node4 +nodes +extent +sex, data=colon.death1)
d.model10.6.7.2<-coxph(Surv(time, status) ~node4 +nodes +extent +age, data=colon.death1)
```

```
d.model10.6.7.3<-coxph(Surv(time, status) ~node4 +nodes +extent +obstruct, data=colon.death1)
d.model10.6.7.4<-coxph(Surv(time, status) ~node4 +nodes +extent +perfor, data=colon.death1)
d.model10.6.7.5<-coxph(Surv(time, status) ~node4 +nodes +extent +adhere, data=colon.death1)
d.model10.6.7.6<-coxph(Surv(time, status) ~node4 +nodes +extent +differ, data=colon.death1)
d.model10.6.7.7<-coxph(Surv(time, status) ~node4 +nodes +extent +surg, data=colon.death1)

BIC(d.model10.6.7.1, d.model10.6.7.2, d.model10.6.7.3, d.model10.6.7.4, d.model10.6.7.5,
d.model10.6.7.6, d.model10.6.7.7)
##                df      BIC
## d.model10.6.7.1  6 5460.479
## d.model10.6.7.2  6 5457.601
## d.model10.6.7.3  6 5456.224
## d.model10.6.7.4  6 5460.276
## d.model10.6.7.5  6 5457.446
## d.model10.6.7.6  7 5462.023
## d.model10.6.7.7  6 5454.236
```

The model with the smallest BIC value in this step is d.model10.6.7.7 which contains covariate node4, nodes, extent and surg.

```
d.model10.6.7.7.1<-coxph(Surv(time, status) ~node4 +nodes +extent +surg +sex, data=colon.death1)
d.model10.6.7.7.2<-coxph(Surv(time, status) ~node4 +nodes +extent +surg +age, data=colon.death1)
d.model10.6.7.7.3<-coxph(Surv(time, status) ~node4 +nodes +extent +surg +obstruct, data=colon.death1)
d.model10.6.7.7.4<-coxph(Surv(time, status) ~node4 +nodes +extent +surg +perfor, data=colon.death1)
d.model10.6.7.7.5<-coxph(Surv(time, status) ~node4 +nodes +extent +surg +adhere, data=colon.death1)
d.model10.6.7.7.6<-coxph(Surv(time, status) ~node4 +nodes +extent +surg +differ, data=colon.death1)

BIC(d.model10.6.7.7.1, d.model10.6.7.7.2, d.model10.6.7.7.3, d.model10.6.7.7.4, d.model10.6.7.7.5,
d.model10.6.7.7.6)
##                  df      BIC
## d.model10.6.7.7.1  7 5460.298
## d.model10.6.7.7.2  7 5457.623
## d.model10.6.7.7.3  7 5456.101
## d.model10.6.7.7.4  7 5460.125
## d.model10.6.7.7.5  7 5457.500
## d.model10.6.7.7.6  8 5461.947
```

In this step, BIC values for all the models are larger than BIC values for all the models in the previous step. Hence, we stopped fitting the model with more covariates.

```
d.model.full <- coxph(Surv(time, status) ~sex +age +obstruct +perfor +adhere +nodes +differ
+extent +surg +node4, data=colon.death1)

BIC(d.model.full, d.model10.6.7.7, d.model10.6.7, d.model10.6, d.model10)
##                df      BIC
## d.model.full   13 5482.617
## d.model10.6.7.7  6 5454.236
## d.model10.6.7    5 5454.430
## d.model10.6      2 5455.527
```

```
## d.model10        1 5458.201
```

We fit a model with all the covariates in the dataset to obtain its BIC value. Then we compare its BIC value to the smallest BIC value of each previous step to obtain the best model.

The resulting model with the lowest BIC is:

Surv(time, status) ~node4 +nodes +extent +surg

Next, we used the Analysis of Deviance procedure to get the proper Likelihood Ratio Test to confirm if each of the covariates selected by the forward selection method is significant to include in the Cox Proportional Model.

```
anova(d.model10.6.7.7)
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##             loglik   Chisq    Df    Pr(>|Chi|)
## NULL  -2767.9
## node4 -2726.1 83.6852    1  < 2.2e-16 ***
## nodes -2721.7  8.7382    1   0.0031162 **
## extent -2712.1 19.2886  3   0.0002383 ***
## surg    -2708.9  6.2574    1    0.0123674 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that p-value for each covariate is smaller than 0.05 which means node4, extent, surg and nodes have a significant effect on time until death. Therefore, we will include these four covariates in our Cox PH model.

Interaction Term

Since node4 is characterized as 1 when the number of nodes is greater 4 and 0 when the number of nodes is less than 4, we believe there may be an interaction between covariates nodes and node4. Therefore, we perform a likelihood ratio test to see if the interaction between nodes and node4 should be included in the model

```
#interaction between node4 and nodes
death.interaction <-coxph(Surv(time, status)~node4 +extent+surg+nodes+rx+node4*nodes,
                  data=colon.death1)
anova(death.interaction)
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##                 loglik   Chisq Df Pr(>|Chi|)
## NULL          -2767.9
```

```
## node4          -2726.1 83.6852  1  < 2.2e-16 ***
## extent         -2715.7 20.7944  3  0.0001162 ***
## surg           -2712.8  5.6344  1  0.0176115 *
## nodes          -2708.9  7.8554  1  0.0050670 **
## rx             -2703.9 10.0117  2  0.0066987 **
## node4:nodes -2702.1  3.6855  1  0.0548874 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value in this test is greater than 0.05, so it's not significant. Thus, we decide not to include the interaction in our cox proportional model.

Model evaluation

C-log-log Plot

To check the proportional hazards assumption for this model, we use a diagnostic plot such as the c-log-log plot. Before we plot the C-log-log plot, we need to account for the observations we previously omitted from the data. Hence, we created a second death dataset colon.death2 and replaced the NA values in covariate nodes with its mean and conditioned the NA values in covariate differ into an additional factor level. Then we checked the significance of each covariate again using Analysis of Deviance procedure to ensure that our previous model is still valid.

```
colon.death2 <-colon.death
colon.death2$nodes[is.na(colon.death2$nodes)] <-mean(colon.death2$nodes, na.rm =TRUE)
colon.death2$differ <-factor(colon.death2$differ, exclude=NULL)
d.model.NA.10.6.7.7<-coxph(Surv(time, status) ~node4 +nodes +extent +surg, data=colon.death2)

anova(d.model.NA.10.6.7.7)
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##        loglik   Chisq Df Pr(>|Chi|)
## NULL  -2930.2
## node4 -2885.7 88.9535  1  < 2.2e-16 ***
## nodes -2880.9  9.5643  1 0.0019840 **
## extent -2870.8 20.2325  3  0.0001519 ***
## surg  -2867.2  7.2544  1 0.0070728 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After we checked that all the covariates were significant, we proceeded with plotting the C-log-log plot for each covariate we found to be significant, starting with the covariate node4.

```
dnode4.fit <-survfit(Surv(time, status) ~node4, data =colon.death2)
ggsurvplot(dnode4.fit , conf.int =F,
        fun ="cloglog",
        xlim =c(20, 5000),
        title ="C-Log-Log for Colon Cancer Mortality \nby node4",
```

```
        xlab ="Time (until death) \nin Days")
```

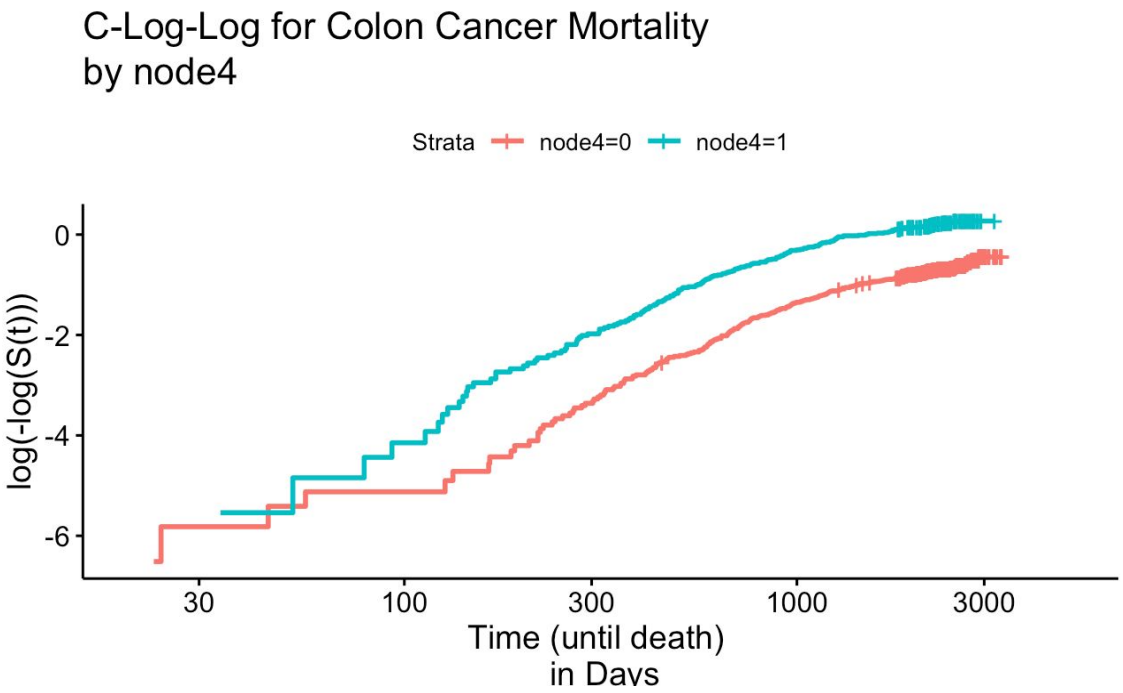

Figure 2: C-log-log plot for covariate node4 in the mortality model

In figure 2, the two curves in this C-log-log plot cross over at the beginning of the study but appear to be parallel to each other after 100 days. Since the data is oftentimes noisy at the beginning of the study, the cross over does not cause too much concern. Overall, we believe that the cox proportional assumption is appropriate for the covariate node4 since the curves are consistently parallel throughout most of the study.

We continue to plot the C-log-log plot for the covariate extent.

```
dextent.fit <-survfit(Surv(time, status) ~extent, data =colon.death2)
ggsurvplot(dextent.fit , conf.int =F,
        fun ="cloglog",
        xlim =c(20, 5000),
        title ="C-Log-Log for Colon Cancer Mortality \nby extent",
        xlab ="Time (until death) \nin Days")
```

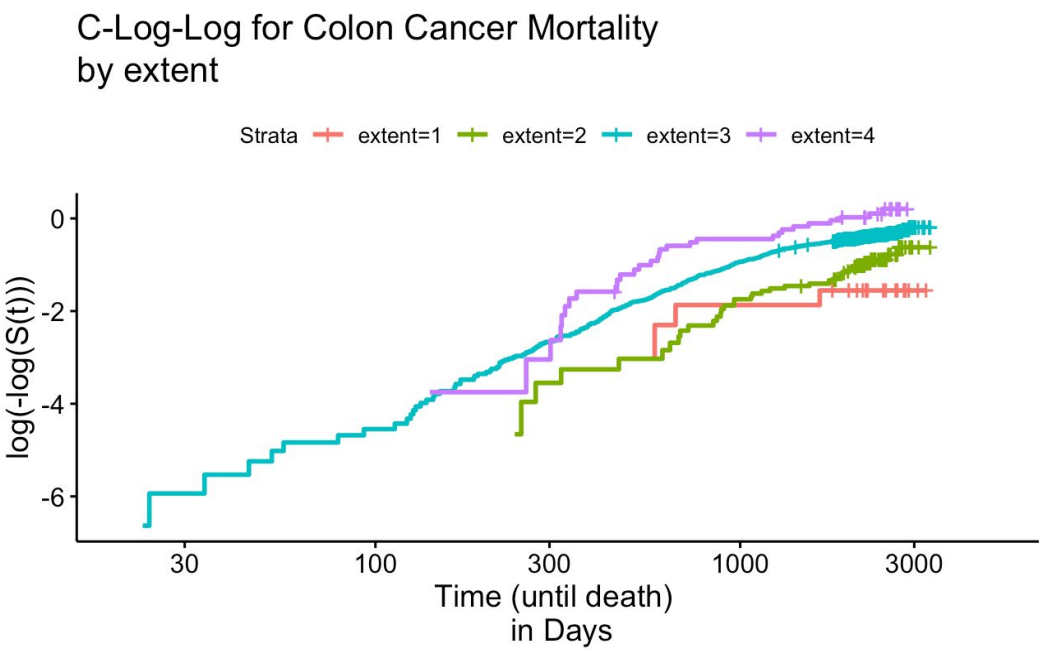

Figure 3: C-log-log plot for covariate extent in the mortality model

In figure 3, the curves in the C-log-log plot are crossing over after 100 days. Since there are not enough data points in each extent group to show a more comprehensive trend, it's hard for us to make a decision based on the plot.

Next we proceed by plotting the C-log-log for the covariate surg.

```
dsurg.fit <-survfit(Surv(time, status) ~surg, data =colon.death2)
ggsurvplot(dsurg.fit , conf.int =F,
        fun ="cloglog",
        xlim =c(20, 5000),
        title ="C-Log-Log for Colon Cancer Mortality \nby surg",
        xlab ="Time (until death) \nin Days")
```

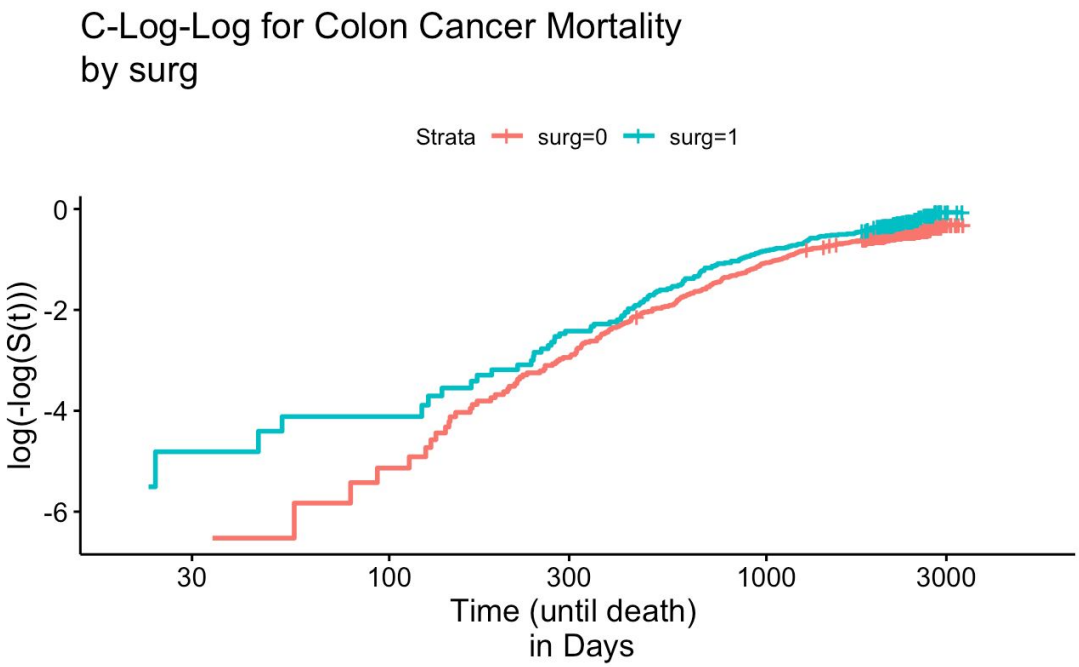

Figure 4: C-log-log plot for covariate surg in the mortality model

In the C-log-log plot for covariate surg, the distance between the two curves begin to narrow after 100 days which causes some concern that the assumption might be violated. However, since the curves seem to be parallel thereafter, we decide that the assumption is appropriate to use for this covariate.

Lastly, we plotted the C-log-log for the covariate rx.

```
drx.fit <- survfit(Surv(time, status) ~ rx, data = colon.death2)

ggsurvplot(drx.fit , conf.int = F,
           fun = "cloglog",
           xlim = c(20, 5000),
           title = "C-Log-Log for Colon Cancer Mortality \nby surg",
           xlab = "Time (until death) \n in Days")
```

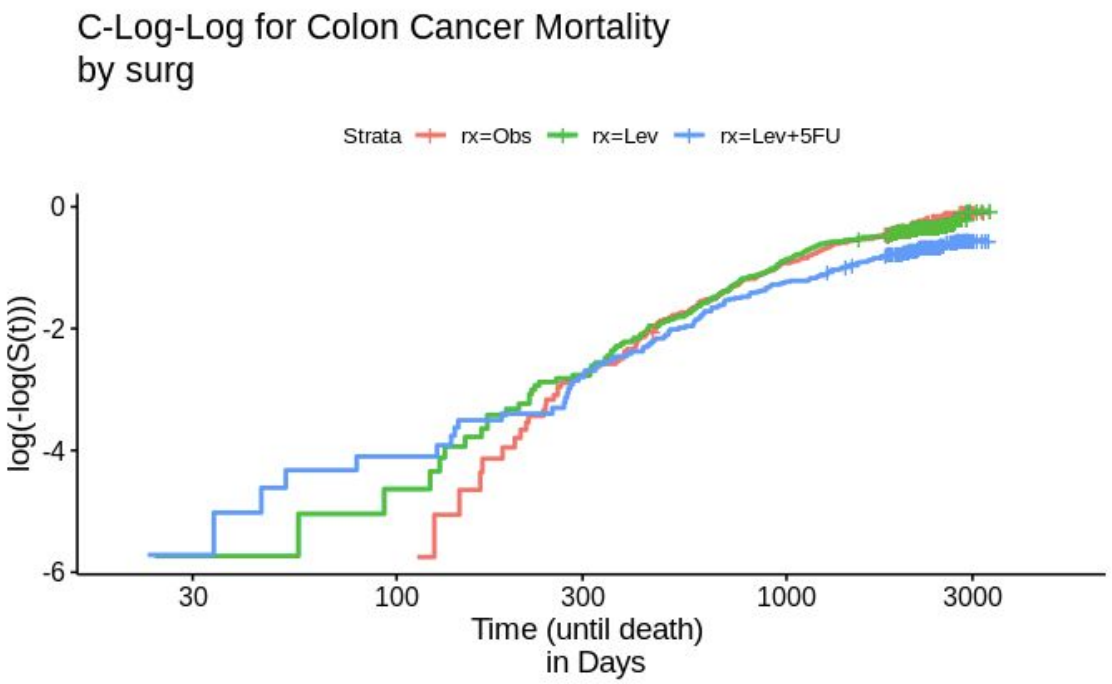

Figure 5: C-log-log plot for covariate surg in the mortality model

Similarly in Figure 5, the distance between three treatment curves begin to narrow after 300 days which causes some concern that the assumption might be violated. However, it is also reasonable to assume that the curves are wider apparent earlier in the study since there are less occurrences of death before 300 days. Hence we believe it is best to ignore the noisiness of the plot since the curves are roughly parallel after 300 days. Thus, the cox proportional hazards assumption is valid to use for this covariate.

Goodness of Fit Test

We used the cox.zph function to further justify our conclusion.

```
cox.zph(coxph(Surv(time, status) ~ node4 + nodes + extent + surg + rx, data=colon.death2))
##          chisq df          p
## node4   5.4233  1 0.020
## nodes   0.3298  1 0.566
## extent  7.2640  3 0.064
## surg 0.0254  1 0.873
## rx     2.2742  2 0.321
## GLOBAL 16.8941  8 0.031
```

Since the p value for node4 is less than 0.05, there is significant evidence that the cox proportional model assumption is violated for variable node4. However, we observed in Figure 2 that the C-log-log curves for covariate node4 appear parallel. We believe that the significance of p-value might be due to some noise in the data at the beginning of the study. All in all, we

conclude that the cox proportional hazards assumption has been met and is reasonable to use for this model.

The Final Model

With the inclusion of treatment, our final model is given by:

Surv(time, status) ~node4 +nodes +extent +surg+rx

```
d.final.model <-coxph(Surv(time, status) ~node4 +nodes +extent +surg +rx, data=colon.death2)
summary(d.final.model)
## Call:
## coxph(formula = Surv(time, status) ~ node4 + nodes + extent +
##     surg + rx, data = colon.death2)
##
##   n= 929, number of events= 452
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## node41       0.63953   1.89560  0.13342 4.793 1.64e-06 ***
## nodes        0.04756   1.04871  0.01439 3.306 0.000948 ***
## extent2      0.37759   1.45877  0.52839 0.715 0.474846
## extent3      0.88594   2.42526  0.50470 1.755 0.079192 .
## extent4      1.37799   3.96692  0.53515 2.575 0.010024 *
## surg1        0.26541   1.30397  0.10340 2.567 0.010260 *
## rxLev       -0.04132   0.95952  0.11087 -0.373 0.709369
## rxLev+5FU -0.36859   0.69171 0.11917 -3.093 0.001982 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## node41         1.8956     0.5275    1.4594   2.4622
## nodes          1.0487     0.9535    1.0196   1.0787
## extent2        1.4588     0.6855    0.5179   4.1091
## extent3        2.4253     0.4123    0.9019   6.5216
## extent4        3.9669     0.2521    1.3898  11.3232
## surg1          1.3040     0.7669    1.0648   1.5969
## rxLev          0.9595     1.0422    0.7721   1.1924
## rxLev+5FU    0.6917     1.4457    0.5476   0.8737
##
## Concordance= 0.665  (se = 0.013 )
## Likelihood ratio test= 137.5  on 8 df,  p=<2e-16
## Wald test            = 146.8  on 8 df,  p=<2e-16
## Score (logrank) test = 159.4  on 8 df,  p=<2e-16
```

From the results of the summary function for the final mortality model, we see that after we control for all other covariates in the model the coefficient for covariate rxLev is -0.041 with a p-value of 0.70, and the hazard ratio between the group treated with rxLev and the observation group is 0.96. Furthermore, the confidence interval for the hazard ratio is (0.77, 1.19).

On the other hand, the coefficient for covariate rxLev is -0.37 with a p-value of 0.0020, and the hazard ratio between the group treated with rexLev+5Fu and the observation group is 0.69. Furthermore, the confidence interval for the hazard ratio is (0.55, 0.87).

Research Question 1

We can now answer our first research question which asks whether treatments Levamisole and Levamisole+5-FU help improve the survival rate in colon cancer patients by estimating the survival probability for different treatment groups. Our result indicates that the hazard rate for the group treated with Levamisole is 96% of what it is for the observation group. Since the p-value for the coefficient is 0.70 which is greater than 0.05, there is no significant evidence to indicate a difference between the hazard rate of the group treated with Levamisole and the hazard rate of the observation group. Therefore, we conclude that the treatment Levamisole is not effective on improving the survival probability in colon cancer patients.

On the other hand, our result indicates that the hazard rate for the group treated with Levamisole+5-FU is 69% of the hazard rate for the observation group. Since the p-value for the coefficient is close to 0, this indicates that there is a significant difference between the hazard rate of the group treated with Levamisole+5-FU and the hazard rate for the observation group. Additionally, since the coefficient is negative then this means that the hazard rate for patients treated with Levamisole+5-FU is in fact less than the hazard rate of the observed group.

Therefore, we conclude that the treatment Levamisole+5-FU is effective on improving the survival rate in colon cancer patients. There is roughly a 30% chance that patients who received the treatment Levamisole+5-FU survive longer than patients who do not receive any treatments and patients who received Levamisole alone.

Survival Analysis for the Event of Recurrence

Model Building Process

Examining the data

In order to answer our second research question which is to study the time until an recurrence event, we create a marginal model for recurrence by subsetting the colon data to only include observations with *etype* 1. To get an overview of the recurrence subset we use the survfit function and plot the Kaplan-Meier Estimate between the three different treatments.

```
#subset recurrence data
colon.recurrence <-subset(colon, etype ==1)

#plot KM estimate for recurrence data
recurrence.fit <-survfit(Surv(time, status) ~ rx, data = colon.recurrence)

ggsurvplot(recurrence.fit, conf.int=F,
           title ='Kaplan-Meier Curve for Colon Cancer \nRecurrence by Treatment',
           xlab='Times (until death) \nin Days')
```

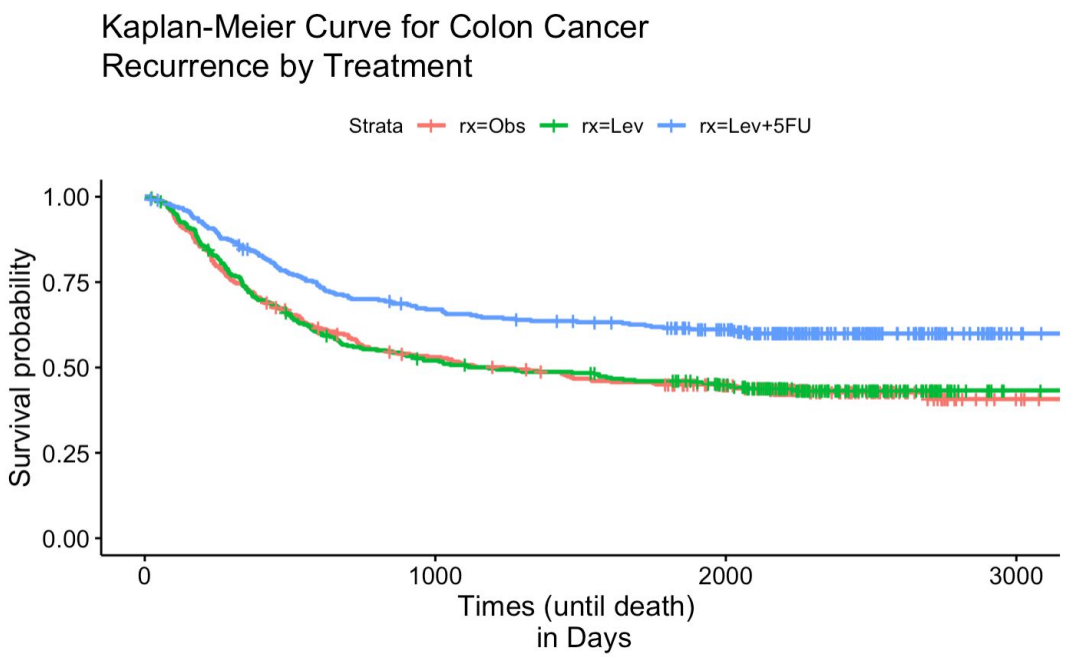

Figure 6: Recurrence probability for colon cancer patients based on treatment.

From figure 6, we see that the survival probability appears to be higher for patients who received the adjuvant treatment with levamisole plus fluorouracil (Lev+5Fu) than patients who received no further treatment and patients who received the treatment with levamisole alone.

Moreover, similar to our survival analysis for the event of death, we removed the NA values before we proceed with the variable selection process.

```
#removing NA values
colon.recurrence1 <- na.omit(colon.recurrence)
```

Variable Selection

Similar to our previous variable selection process, we want to determine which variables have an effect on time to recurrence by using the forward selection method with Bayesian Information Criterion (BIC).

```
r.model1 <-coxph(Surv(time, status) ~sex, data=colon.recurrence1)
r.model2 <-coxph(Surv(time, status) ~age, data=colon.recurrence1)
r.model3 <-coxph(Surv(time, status) ~obstruct, data=colon.recurrence1)
r.model4 <-coxph(Surv(time, status) ~perfor, data=colon.recurrence1)
r.model5 <-coxph(Surv(time, status) ~adhere, data=colon.recurrence1)
r.model6 <-coxph(Surv(time, status) ~nodes, data=colon.recurrence1)
r.model7 <-coxph(Surv(time, status) ~differ, data=colon.recurrence1)
r.model8 <-coxph(Surv(time, status) ~extent, data=colon.recurrence1)
r.model9 <-coxph(Surv(time, status) ~surg, data=colon.recurrence1)
r.model10 <-coxph(Surv(time, status) ~node4, data=colon.recurrence1)
```

```
BIC(r.model1, r.model2,r.model3, r.model4,r.model5,r.model6,r.model7,r.model8,r.model9,r.model10)
##           df     BIC
## r.model1   1 5758.946
## r.model2   1 5758.128
## r.model3   1 5756.859
## r.model4   1 5758.395
## r.model5   1 5755.234
## r.model6   1 5698.941
## r.model7   2 5756.311
## r.model8   3 5744.420
## r.model9   1 5755.739
## r.model10  1 5689.597
```

The model with the smallest BIC value in this step is d.model10 which contains covariate node4.

```
r.model10.1<-coxph(Surv(time, status) ~node4+sex, data=colon.recurrence1)
r.model10.2<-coxph(Surv(time, status) ~node4+age, data=colon.recurrence1)
r.model10.3<-coxph(Surv(time, status) ~node4+obstruct, data=colon.recurrence1)
r.model10.4<-coxph(Surv(time, status) ~node4+perfor, data=colon.recurrence1)
r.model10.5<-coxph(Surv(time, status) ~node4+adhere, data=colon.recurrence1)
r.model10.6<-coxph(Surv(time, status) ~node4+differ, data=colon.recurrence1)
r.model10.7<-coxph(Surv(time, status) ~node4+extent, data=colon.recurrence1)
r.model10.8<-coxph(Surv(time, status) ~node4+surg, data=colon.recurrence1)
r.model10.9<-coxph(Surv(time, status) ~node4+nodes, data=colon.recurrence1)

BIC(r.model10.1,r.model10.2,r.model10.3,r.model10.4,r.model10.5,r.model10.6,r.model10.7,r.model10.8
,r.model10.9)
##            df     BIC
## r.model10.1 2 5694.438
## r.model10.2 2 5695.111
## r.model10.3 2 5690.740
## r.model10.4 2 5693.106
## r.model10.5 2 5691.062
## r.model10.6 3 5695.820
## r.model10.7 4 5685.235
## r.model10.8 2 5689.928
## r.model10.9 2 5687.582
```

The model with the smallest BIC value in this step is d.model10.7 which contains covariate node4 and extent.

```
r.model10.7.1<-coxph(Surv(time, status) ~node4+extent+sex, data=colon.recurrence1)
r.model10.7.2<-coxph(Surv(time, status) ~node4+extent+age, data=colon.recurrence1)
r.model10.7.3<-coxph(Surv(time, status) ~node4+extent+obstruct, data=colon.recurrence1)
r.model10.7.4<-coxph(Surv(time, status) ~node4+extent+perfor, data=colon.recurrence1)
r.model10.7.5<-coxph(Surv(time, status) ~node4+extent+adhere, data=colon.recurrence1)
r.model10.7.6<-coxph(Surv(time, status) ~node4+extent+differ, data=colon.recurrence1)
```

```
r.model10.7.7<-coxph(Surv(time, status) ~node4+extent+nodes, data=colon.recurrence1)
r.model10.7.8<-coxph(Surv(time, status) ~node4+extent+surg, data=colon.recurrence1)

BIC(r.model10.7.1,r.model10.7.2,r.model10.7.3,
r.model10.7.4,r.model10.7.5,r.model10.7.6,r.model10.7.7,r.model10.7.8)
##              df      BIC
## r.model10.7.1  5 5690.331
## r.model10.7.2  5 5690.806
## r.model10.7.3  5 5688.145
## r.model10.7.4  5 5689.742
## r.model10.7.5  5 5688.623
## r.model10.7.6  6 5692.914
## r.model10.7.7  5 5685.487
## r.model10.7.8  5 5685.198
```

The model with the smallest BIC value in this step is d.model10.7.8 which contains covariate node4, extent and surg.

```
r.model10.7.8.1<-coxph(Surv(time, status) ~node4+extent+surg+sex, data=colon.recurrence1)
r.model10.7.8.2<-coxph(Surv(time, status) ~node4+extent+surg+age, data=colon.recurrence1)
r.model10.7.8.3<-coxph(Surv(time, status) ~node4+extent+surg+obstruct, data=colon.recurrence1)
r.model10.7.8.4<-coxph(Surv(time, status) ~node4+extent+surg+perfor, data=colon.recurrence1)
r.model10.7.8.5<-coxph(Surv(time, status) ~node4+extent+surg+adhere, data=colon.recurrence1)
r.model10.7.8.6<-coxph(Surv(time, status) ~node4+extent+surg+differ, data=colon.recurrence1)
r.model10.7.8.7<-coxph(Surv(time, status) ~node4+extent+surg+nodes, data=colon.recurrence1)
BIC(r.model10.7.8.1, r.model10.7.8.2,r.model10.7.8.3,
r.model10.7.8.4,r.model10.7.8.5,r.model10.7.8.6,r.model10.7.8.7)
##                df      BIC
## r.model10.7.8.1  6 5690.164
## r.model10.7.8.2  6 5690.688
## r.model10.7.8.3  6 5688.372
## r.model10.7.8.4  6 5689.722
## r.model10.7.8.5  6 5688.796
## r.model10.7.8.6  7 5693.056
## r.model10.7.8.7  6 5685.175
```

The model with the smallest BIC value in this step is d.model10.7.8.7 which contains covariate node4, extent, surg and nodes.

```
r.model10.7.8.7.1<-coxph(Surv(time, status) ~node4+extent+surg+nodes+sex, data=colon.recurrence1)
r.model10.7.8.7.2<-coxph(Surv(time, status) ~node4+extent+surg+nodes+age, data=colon.recurrence1)
r.model10.7.8.7.3<-coxph(Surv(time, status) ~node4+extent+surg+nodes+obstruct,
data=colon.recurrence1)
r.model10.7.8.7.4<-coxph(Surv(time, status) ~node4+extent+surg+nodes+perfor,
data=colon.recurrence1)
r.model10.7.8.7.5<-coxph(Surv(time, status) ~node4+extent+surg+nodes+adhere,
data=colon.recurrence1)
```

```
r.model10.7.8.7.6<-coxph(Surv(time, status) ~node4+extent+surg+nodes+differ,
data=colon.recurrence1)

BIC(r.model10.7.8.7.1,r.model10.7.8.7.2,r.model10.7.8.7.3,
r.model10.7.8.7.4,r.model10.7.8.7.5,r.model10.7.8.7.6)
##                  df      BIC
## r.model10.7.8.7.1  7 5689.832
## r.model10.7.8.7.2  7 5690.669
## r.model10.7.8.7.3  7 5687.991
## r.model10.7.8.7.4  7 5689.794
## r.model10.7.8.7.5  7 5688.849
## r.model10.7.8.7.6  8 5693.684
```

In this step, BIC values for all the models are larger than BIC values for all the models in the previous step. Hence, we stopped fitting the model with more covariates.

```
BIC(r.model10.7.8.7.3,r.model10.7.8.7,r.model10.7.8,r.model10.7,r.model10)
##                       df      BIC
## r.model10.7.8.7.3  7    5687.991
## r.model10.7.8.7    6    5685.175
## r.model10.7.8      5    5685.198
## r.model10.7        4    5685.235
## r.model10          1    5689.597
```

Of all the best models from each step, r.model10.7.8.7 has the smallest BIC value. r.model10.7.8.7 includes covariates node4, nodes, surg and extent.

We used the Analysis of Deviance procedure to get the proper Likelihood Ratio Test to see if each of the covariates selected by the forward selection method is significant to include in the Cox Proportional Model.

```
anova(r.model10.7.8.7)
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##           loglik     Chisq      Df    Pr(>|Chi|)
## NULL  -2877.1
## node4  -2841.8   70.7685   1     < 2.2e-16 ***
## extent  -2830.4   22.6631   3    4.747e-05 ***
## surg     -2827.3   6.1378     1    0.01323 *
## nodes   -2824.3   6.1227     1    0.01335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that p-value for each covariate is smaller than 0.05 which means node4, extent, surg and nodes have the effect on time until recurrence. Therefore, we decide to include these four covariates in our model.

Model evaluation

C-log-log Plot

Next, we want to check whether the cox proportional hazards assumption is appropriate to use for each covariate by plotting the C-log-log plots. Similar to our procedure in analyzing the mortality data, we need to account for the observations we previously omitted from the data. Hence, we created a second recurrence dataset colon.recurrence2 and replaced the NA values in covariate nodes with its mean and conditioned the NA values in covariate differ into an additional factor level. Then we checked the significance of each covariate again using Analysis of Deviance procedure to ensure that our previous model is still valid.

```
#replacing NA values
colon.recurrence2 <-colon.recurrence
colon.recurrence2$nodes[is.na(colon.recurrence2$nodes)] <-mean(colon.recurrence2$nodes,
na.rm =TRUE)
colon.recurrence2$differ[is.na(colon.recurrence2$differ)] <-2
colon.recurrence2$differ <-factor(colon.recurrence2$differ, exclude=NULL)
```

```
plot(survfit(recurrence2.surv~node4, data=colon.recurrence2),
    fun='cloglog',
    col =c(2,3),
    ylab='log-log(S(t))',
    xlab='Time(days)',
    main="C-log-log plot for Covariate node4")
legend("topleft",legend =c("more than 4 positive nodes","less than 4 nodes"), col=c(2,3),lty=1)
```

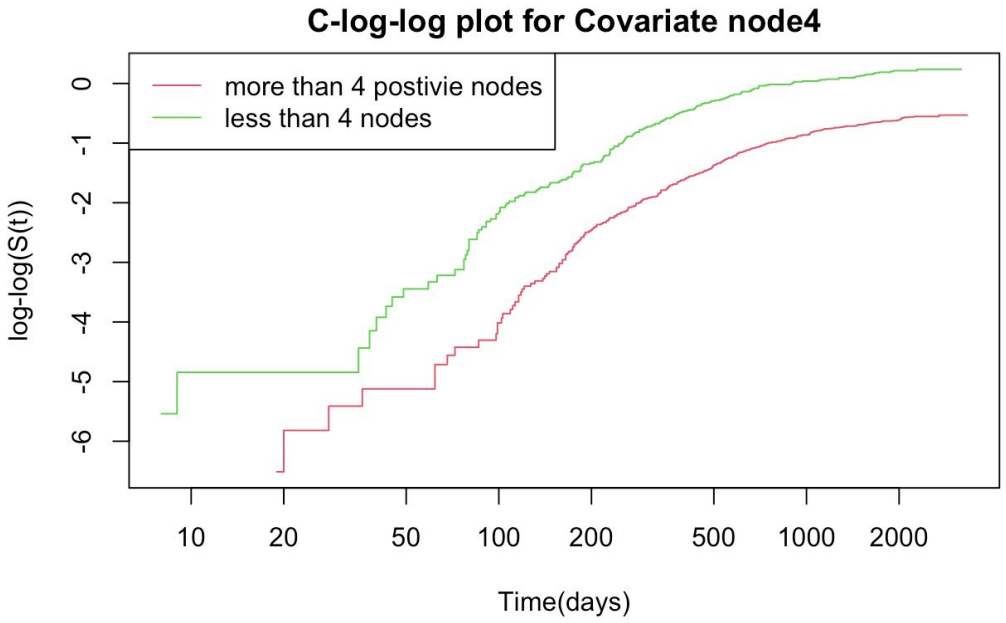

Figure 7: C-log-log plot for covariate node4 in the recurrence model

In figure 7, the two curves in this C-log-log plot appear to be parallel to each other, so we think the cox proportional assumption is approritate for the covariate node4.

```
plot(survfit(recurrence2.surv~extent, data=colon.recurrence2),
    fun='cloglog',
    col =c(2,3,4,5),
    ylab='log-log(S(t))',
    xlab='Time(days)',
    main="C-log-log plot for Covariate extent")
legend("topleft",legend =c("1-submucosa","2-muscle","3-serosa", "4-contiguous structures"),
col=c(2,3,4,5),lty=1)
```

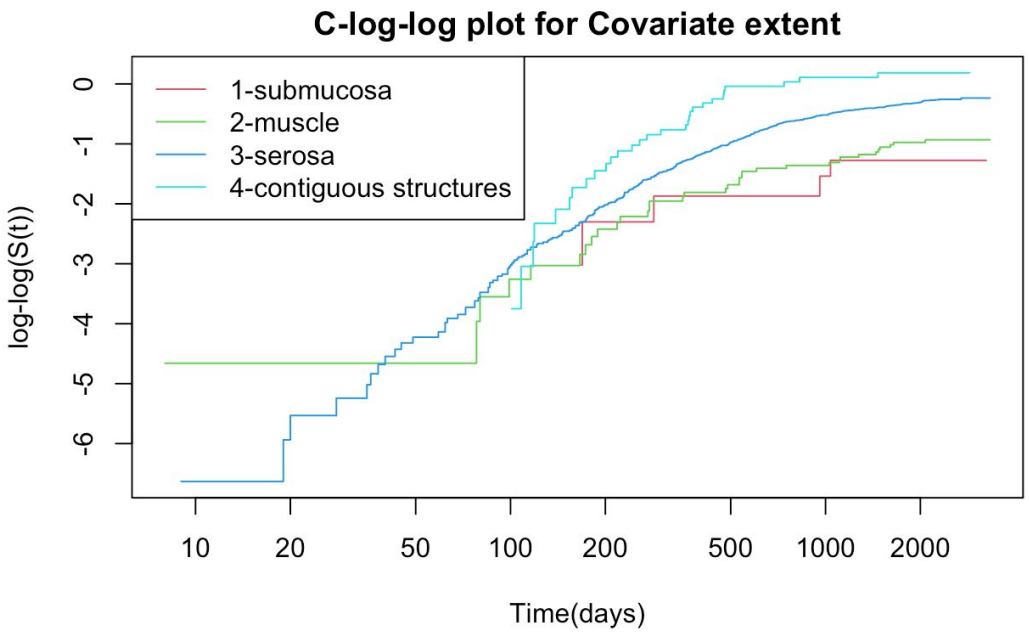

Figure 8: C-log-log plot for covariate extent in the recurrence model

In figure 8, the curves in the C-log-log plot are crossing over after 100 days. Since there are not enough data points in each group to show a more comprehensive trend, it's hard for us to make a decision based on the plot.

```
plot(survfit(recurrence2.surv~surg, data=colon.recurrence2),
    fun='cloglog',
    col =c(2,3),
    ylab='log-log(S(t))',
    xlab='Time(days)',
    main="C-log-log plot for Covariate surg")
legend("topleft",legend =c("short","long"), col=c(2,3),lty=1)
```

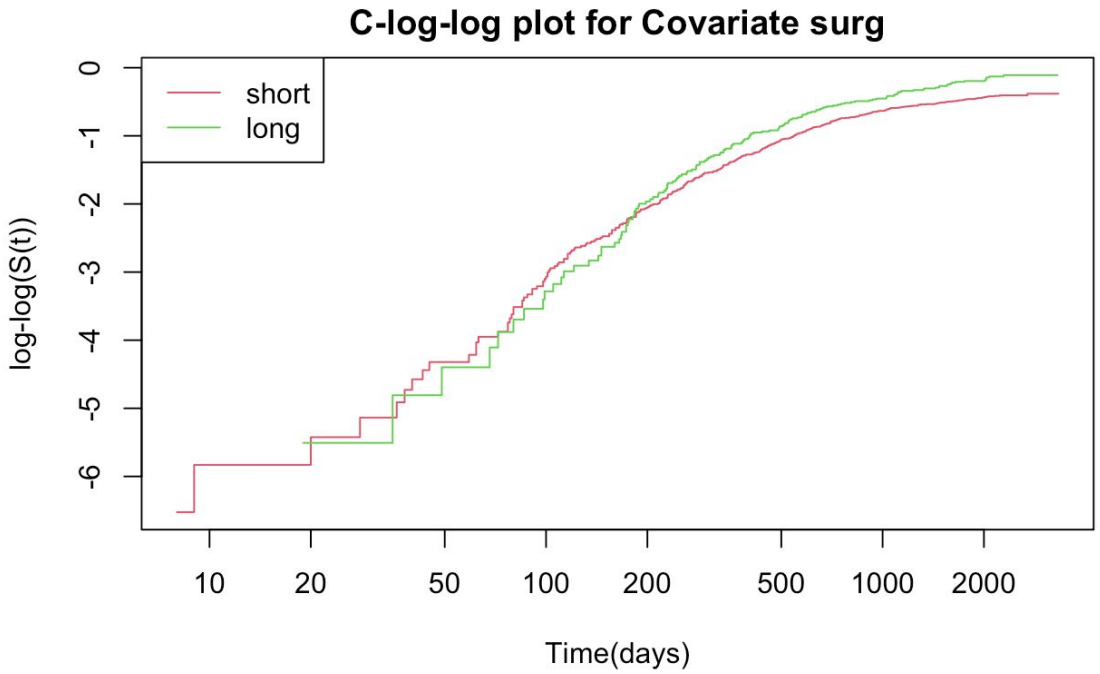

Figure 9: C-log-log plot for covariate surg in the recurrence model

The two curves in the C-log-log plot for covariate surg are crossing over and very close to each other. We conclude that the assumption is appropriate to use for this covariate because crossing over means the two groups have the same hazard ratio.

```
plot(survfit(recurrence2.surv~rx, data=colon.recurrence2),
    fun='cloglog',
    col =c(2,3,4),
    ylab='log-log(S(t))',
    xlab='Time(days)',
    main="C-log-log plot for Covariate rx")
legend("topleft",legend =c("Obs","Lev","Lev+5FU"), col=c(2,3,4),lty=1)
```

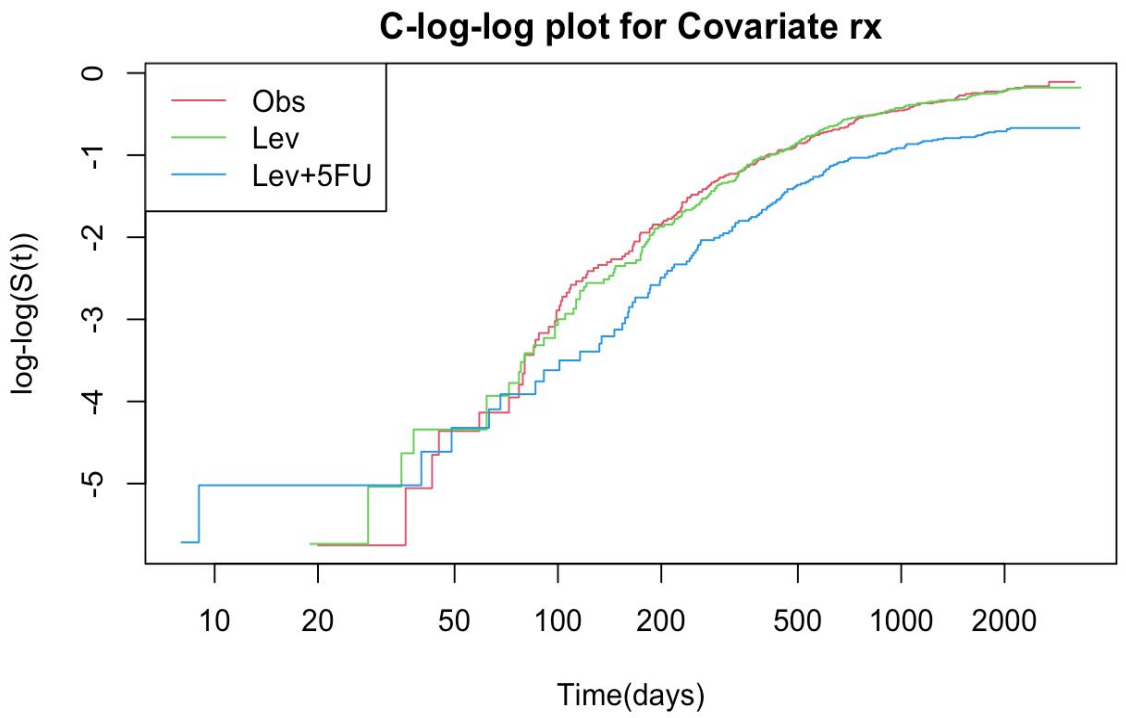

Figure 10: C-log-log plot for covariate rx in the recurrence model

In figure 10, the three curves in the C-log-log plot for covariate rx are overlapping before 100 days, but after 100 days the curves for Obs and Lev are crossing over and the two curves appear to be parallel to the curve for Lev+5Fu. We are somewhat concerned that the assumption might be violated because of the crossover before 100 days.

Goodness of Fit Test

To justify our conclusion drawn from the previous C-log-log plots, we used the cox.zph() function.

```
cox.zph(coxph(formula =Surv(time, status) ~node4 +extent+surg+nodes+rx, data =colon.recurrence2))
```

```
##                chisq  df      p
## node4        10.3190   1 0.0013
## extent        1.5094   3 0.6801
## surg          1.2445   1 0.2646
## nodes         1.8314   1 0.1760
## rx            0.0672   2 0.9670
## GLOBAL       15.0548   8 0.0581
```

Since the p value for node4 is less than 0.05, there is significant evidence that the cox proportional model assumption is violated for variable node4. However, the C-log-log plot for covariate node4 looks parallel, so we think the significance of p-value might be due to some noise in the data at the beginning of the study. At last, we decided the cox proportional model assumption is appropriate use for this model.

The Final Model

The final model for studying time to recurrence is:

Surv(time, status) ~ node4+extent+surg+nodes+rx

```
colon.recurrence.coxph <-coxph(Surv(time, status) ~node4+extent+surg+nodes+rx,
data=colon.recurrence1)
summary(colon.recurrence.coxph)
## Call:
## coxph(formula = Surv(time, status) ~ node4 + extent + surg +
##     nodes + rx, data = colon.recurrence1)
##
##   n= 888, number of events= 446
##
##                 coef     exp(coef)  se(coef)     z        Pr(>|z|)
## node41        0.59457   1.81224   0.14116    4.212   2.53e-05 ***
## extent2       0.28376   1.32811   0.52999    0.535   0.59238
## extent3       0.83970   2.31568   0.50469    1.664   0.09615 .
## extent4       1.34753   3.84790   0.54182    2.487   0.01288 *
## surg1         0.23901   1.27000   0.10402    2.298   0.02158 *
## nodes         0.03888   1.03965   0.01498    2.595   0.00946 **
## rxLev        -0.04734   0.95376  0.11055    -0.428   0.66849
## rxLev+5FU  -0.49881   0.60725  0.12170    -4.099    4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## node41         1.8122    0.5518    1.3742   2.3899
## extent2        1.3281    0.7529    0.4700   3.7529
## extent3        2.3157    0.4318   0.8612    6.2268
## extent4        3.8479    0.2599    1.3305  11.1282
## surg1          1.2700    0.7874    1.0358   1.5572
## nodes          1.0396    0.9619    1.0096   1.0706
## rxLev          0.9538    1.0485    0.7680   1.1845
## rxLev+5FU    0.6073    1.6468    0.4784   0.7708
##
## Concordance= 0.661  (se = 0.013 )
## Likelihood ratio test= 126.2  on 8 df,  p=<2e-16
## Wald test            = 132.2  on 8 df,  p=<2e-16
## Score (logrank) test = 141.6  on 8 df,  p=<2e-16
```

From the results of the summary function for the final model, we see that after control for all other covariates in the model the coefficient for rxLev is -0.047 with a p-value of 0.67, and the hazard ratio between the group treated with rxLev and the observation group is 0.95. Furthermore, the confidence interval for the hazard ratio is (0.77, 1.18) which contains 1.

On the other hand, the coefficient for rxLev+5Fu is -0.50 with a p-value of 0, and the hazard ratio between the group treated with rexLev+5Fu is 0.60. Furthermore, the confidence interval for the hazard ratio is (0.48, 0.77) which is below 1.

Research Question 2

We answer our second research question which asks whether treatments Levamisole and Levamisole+5-FU help delay time to recurrence in colon cancer patients by estimating the survival probability for different treatment groups. Our result indicates that the hazard rate for the group treated with Levamisole is 95% of what it is for the observation group. Since the p-value for the coefficient is 0.67 which is greater than 0.05, there is no significant evidence to indicate a difference between the hazard rate of the group treated with Levamisole and the hazard rate of the observation group. Therefore, we conclude that the treatment Levamisole is not effective on delay time to recurrence in colon cancer patients.

On the other hand, our result indicates that the hazard rate for the group treated with Levamisole+5-FU is 60% of the hazard rate for the observation group. Since the p-value for the coefficient is close to 0, this indicates that there is a significant difference between the hazard rate of the group treated with Levamisole+5-FU and the hazard rate for the observation group. Additionally, since the coefficient is negative then this means that the hazard rate for patients treated with Levamisole+5-FU is in fact less than the hazard rate of the observed group.

Therefore, we conclude that the treatment Levamisole+5-FU is effective on delaying time to recurrence in colon cancer patients. The chance for patients who received the treatment Levamisole+5-FU to have a recurrence event is about 40% less than patients who do not receive any treatments.

Survival Analysis for Event of Mortality in Gap Model

Examining the data

For our third and last research question we wish to determine the effectiveness of the treatments on the survival rate after the event of a recurrence. To do so, we create a new colon data set with an additional column for start time of recurrence for each subject. We also remove the observations in which there was no event of recurrent nor an event of death. To get an overview of the gap subset we use the survfit function and plot the Kaplan-Meier Estimate between the three different treatments.

```
#sorting the colon data
new.colon <- colon %>% arrange(id, time, etype)

#creating a start time variable
new.colon$tstart <- ifelse(colon$etype == 1, colon$time, 0)

colon.gap <- subset(new.colon, new.colon$tstart != new.colon$time & new.colon$etype == 2)

g.death.fit <- survfit(Surv(time - tstart, status) ~ rx, data = colon.gap)

ggsurvplot(g.death.fit, conf.int = F,
```

```
        title = "Kaplan-Meier Curve for Colon Cancer Death \nAfter Event of Reccurence by Treatment",
        xlab = "Time (until death) \n in Days")
```

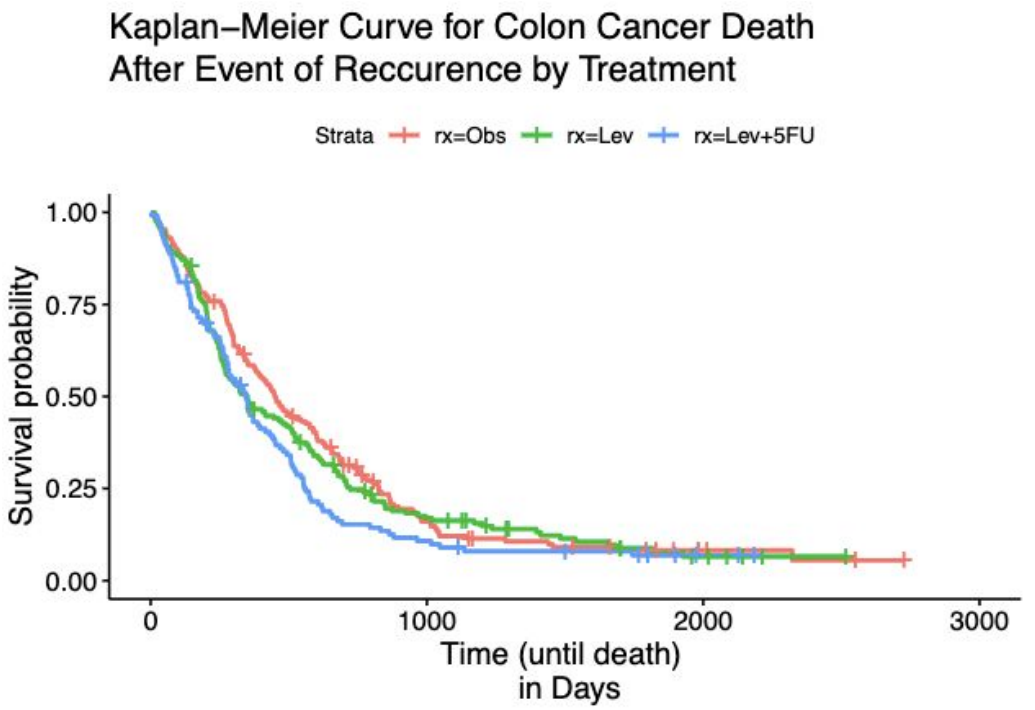

Figure 11: Survival probability for colon cancer patients after a recurrence event based on treatment.

Based on Figure 11, the colon cancer patients within each group appear to have a similar survival probability after the event of recurrence. It is also a possibility that the treatment, Levamisole+5-FU, is not as effective after a recurrence of colon cancer since it has a steeper drop than the other observations.

Moreover, similar to our survival analysis in our marginal models, we removed the NA values before we proceed with the variable selection process.

```
#removing NA values
colon.recurrence1 <- na.omit(colon.recurrence)
```

Variable Selection

We then continue with the forward selection process using Bayesian Information Criterion (BIC) to determine the covariates that best represent an appropriate cox proportional hazards model for the event of death after a recurrence. Within each step, we choose the model that has the lowest BIC value.

```
g.model1 <- coxph(Surv(time - tstart, status) ~ sex, data=colon.gap1)
g.model2 <- coxph(Surv(time - tstart, status) ~ age, data=colon.gap1)
g.model3 <- coxph(Surv(time - tstart, status) ~ obstruct, data=colon.gap1)
g.model4 <- coxph(Surv(time - tstart, status) ~ perfor, data=colon.gap1)
```

```
g.model5 <- coxph(Surv(time - tstart, status) ~ adhere, data=colon.gap1)
g.model6 <- coxph(Surv(time - tstart, status) ~ nodes, data=colon.gap1)
g.model7 <- coxph(Surv(time - tstart, status) ~ differ, data=colon.gap1)
g.model8 <- coxph(Surv(time - tstart, status) ~ extent, data=colon.gap1)
g.model9 <- coxph(Surv(time - tstart, status) ~ surg, data=colon.gap1)
g.model10 <- coxph(Surv(time - tstart, status) ~ node4, data=colon.gap1)

BIC(g.model1, g.model2, g.model3, g.model4, g.model5, g.model6, g.model7, g.model8, g.model9,
g.model10)
##         df        BIC
## g.model1   1 4123.385
## g.model2   1 4122.150
## g.model3   1 4124.603
## g.model4   1 4124.256
## g.model5   1 4125.561
## g.model6   1 4105.940
## g.model7   2 4123.420
## g.model8   3 4130.487
## g.model9   1 4125.822
## g.model10  1 4101.904
```

The model with the smallest BIC value in this step is g.model10 which contains covariate node4.

```
g.model10.1 <- coxph(Surv(time - tstart, status) ~ node4 + sex, data=colon.gap1)
g.model10.2 <- coxph(Surv(time - tstart, status) ~ node4 + age, data=colon.gap1)
g.model10.3 <- coxph(Surv(time - tstart, status) ~ node4 + obstruct, data=colon.gap1)
g.model10.4 <- coxph(Surv(time - tstart, status) ~ node4 + perfor, data=colon.gap1)
g.model10.5 <- coxph(Surv(time - tstart, status) ~ node4 + adhere, data=colon.gap1)
g.model10.6 <- coxph(Surv(time - tstart, status) ~ node4 + nodes, data=colon.gap1)
g.model10.7 <- coxph(Surv(time - tstart, status) ~ node4 + differ, data=colon.gap1)
g.model10.8 <- coxph(Surv(time - tstart, status) ~ node4 + extent, data=colon.gap1)
g.model10.9 <- coxph(Surv(time - tstart, status) ~ node4 + surg, data=colon.gap1)

BIC(g.model10.1, g.model10.2, g.model10.3, g.model10.4, g.model10.5, g.model10.6, g.model10.7,
g.model10.8, g.model10.9)
##         df        BIC
## g.model10.1  2 4102.886
## g.model10.2  2 4100.377
## g.model10.3  2 4105.998
## g.model10.4  2 4106.149
## g.model10.5  2 4106.646
## g.model10.6  2 4105.940
## g.model10.7  3 4109.455
## g.model10.8  4 4113.364
## g.model10.9  2 4107.856
```

The model with the smallest BIC value in this step is d.model10.2 which contains covariate node4 and age.

```
g.model10.2.1 <- coxph(Surv(time - tstart, status) ~ node4 + age + sex, data=colon.gap1)
g.model10.2.3 <- coxph(Surv(time - tstart, status) ~ node4 + age + obstruct, data=colon.gap1)
g.model10.2.4 <- coxph(Surv(time - tstart, status) ~ node4 + age + perfor, data=colon.gap1)
g.model10.2.5 <- coxph(Surv(time - tstart, status) ~ node4 + age + adhere, data=colon.gap1)
g.model10.2.6 <- coxph(Surv(time - tstart, status) ~ node4 + age + nodes, data=colon.gap1)
g.model10.2.2 <- coxph(Surv(time - tstart, status) ~ node4 + age + differ, data=colon.gap1)
g.model10.2.7 <- coxph(Surv(time - tstart, status) ~ node4 + age + extent, data=colon.gap1)
g.model10.2.8 <- coxph(Surv(time - tstart, status) ~ node4 + age + surg, data=colon.gap1)

BIC(g.model10.2.1, g.model10.2.2, g.model10.2.3, g.model10.2.4, g.model10.2.5, g.model10.2.6,
g.model10.2.7, g.model10.2.8)
##                  df      BIC
## g.model10.2.1  3 4101.872
## g.model10.2.2  4 4108.037
## g.model10.2.3  3 4103.612
## g.model10.2.4  3 4104.494
## g.model10.2.5  3 4105.427
## g.model10.2.6  3 4104.561
## g.model10.2.7  5 4112.192
## g.model10.2.8  3 4106.307
```

In this step, BIC values for all the models are larger than BIC values for all the models in the previous step. Hence, we stopped fitting the model with more covariates.

```
g.model.full <- coxph(Surv(time - tstart, status) ~ sex + age + obstruct + perfor + adhere + nodes + differ
+ extent + surg + node4, data=colon.gap1)

BIC(g.model.full, g.model10.2, g.model10)
##                  df      BIC
## g.model.full 13 4142.573
## g.model10.2   2 4100.377
## g.model10     1 4101.904
```

We fit a model with all the covariates in the dataset to obtain its BIC value. Then we compare its BIC value to the smallest BIC value of each previous step to obtain the best model.

The resulting model with the lowest BIC is:

$$\text{Surv(time, status)} \sim \text{node4} + \text{age}$$

Next, we used the Analysis of Deviance procedure to get the proper Likelihood Ratio Test to confirm if each of the covariates selected by the forward selection method is significant to include in the Cox Proportional Model.

```
anova(g.model10.2)
## Analysis of Deviance Table
```

```
##  Cox model: response is Surv(time - tstart, status)
## Terms added sequentially (first to last)
##
##         loglik   Chisq Df Pr(>|Chi|)
## NULL  -2060.1
## node4 -2048.0 24.2323  1  8.539e-07 ***
## age   -2044.2  7.4899  1   0.006205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that p-value for each covariate is smaller than 0.05 which means covariates node4 and age have a significant effect on time until death.

Model evaluation

C-log-log Plot

To check the proportional hazards assumption for this model, we examine the C-log-log plot for both node4 and age. Yet, like we previously did in the marginal models, we first need to account for the observations we previously omitted from the gap data set. Hence, we created a second gap dataset, colon.gap2, and replaced the NA values in covariate nodes with its mean and conditioned the NA values in covariate differ into an additional factor level. Then we checked the significance of each covariate again using Analysis of Deviance procedure to ensure that our previous model is still valid.

```
colon.gap2 <- colon.gap
colon.gap2$nodes[is.na(colon.gap1$nodes)] <- mean(colon.gap1$nodes, na.rm = TRUE)
colon.gap2$differ[is.na(colon.gap1$differ)] <- factor(colon.gap2$differ, exclude=NULL)

g.model.NA.10.2 <- coxph(Surv(time - tstart, status) ~ node4 + age, data=colon.gap2)
anova(g.model.NA.10.2)
## Analysis of Deviance Table
##  Cox model: response is Surv(time - tstart, status)
## Terms added sequentially (first to last)
##
##         loglik   Chisq Df Pr(>|Chi|)
## NULL  -2182.3
## node4 -2170.8 22.9272  1  1.683e-06 ***
## age   -2167.9  5.6902  1     0.01706 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value for node4 and age are less than the significance level, 0.05, we proceeded with plotting the C-log-log plot for each covariate we found to be significant, starting with the covariate node4.

```
gnode4.fit <- survfit(Surv(time - tstart, status) ~ node4, data = colon.gap2)

ggsurvplot(gnode4.fit , conf.int = F,
```

```
        fun = "cloglog",
        xlim = c(20, 5000),
        title = "C-Log-Log for Colon Cancer Mortality After Event of Recurrence \nby node4",
        xlab = "Time (until death) \n in Days")
```

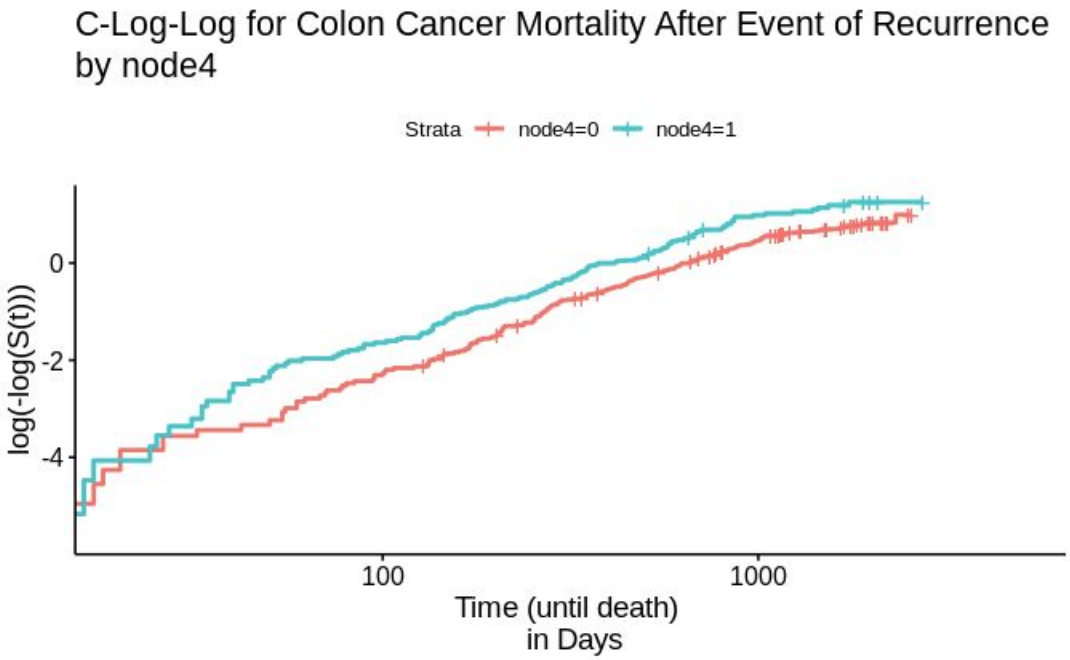

Figure 12: C-log-log plot for covariate node4 in the gap model

In figure 12, the two curves in this C-log-log plot cross over at the beginning of the study but appear to be parallel to each other after 50 days. Since the data is oftentimes noisy at the beginning of the study, the cross over does not cause too much concern. Overall, we believe that the cox proportional assumption is appropriate for the covariate node4 since the curves are consistently parallel throughout most of the study.

We continue to plot the C-log-log plot for the covariate age. To do so, we categorized the patients by age. The first category of patients are those who are between 0 to 55 years old. The second category are patients between the ages 55 and 65. Lastly the older patients are those over the age of 65. Each category has a similar number of observations as shown from the summary table of the colon.gap2 dataset.

```
colon.gap2$age.cat <- cut(colon.gap2$age, breaks = c(0, 55, 65, Inf))

summary(colon.gap2)

##        id         study             rx         sex         age       obstruct
##  Min.   : 1.0   Min.   :1   Obs :174   0:225   Min.   :18.00   0:362
##  1st Qu.:228.0   1st Qu.:1   Lev        :171   1:236   1st Qu.:51.00   1: 99
##  Median :474.0   Median :1   Lev+5FU:116                Median :61.00
##  Mean   :466.1   Mean   :1                              Mean   :58.82
##  3rd Qu.:698.0   3rd Qu.:1                              3rd Qu.:68.00
```

```
##  Max.   :927.0   Max.   :1                       Max.   :85.00
##
##  perfor  adhere         nodes              status             differ    extent  surg
##  0:444   0:380   Min.   : 0.000   Min.   :0.0000   1   : 43   1:  5   0:322
##  1: 17   1: 81   1st Qu.: 2.000   1st Qu.:1.0000   2   :322   2: 34   1:139
##                  Median : 3.000   Median :1.0000   3   : 86   3:393
##                  Mean   : 4.509   Mean   :0.8872   NA's: 10   4: 29
##                  3rd Qu.: 6.000   3rd Qu.:1.0000
##                  Max.   :33.000   Max.   :1.0000
##                  NA's   :11
##  node4         time      etype     tstart              age.cat
##  0:285   Min.   :  34   1:  0   Min.   :   8.0   (0,55]  :157
##  1:176   1st Qu.: 503   2:461   1st Qu.: 204.0   (55,65] :146
##          Median : 887           Median : 384.0   (65,Inf]:158
##          Mean   :1076           Mean   : 541.8
```

After categorizing the patients by age we continued with the C-log-log plot for the categorized covariate age.

```
gage.fit <- survfit(Surv(time - tstart, status) ~ age.cat, data = colon.gap2)

ggsurvplot(gage.fit , conf.int = F,
         fun = "cloglog",
           xlim = c(20, 5000),
           title = "C-Log-Log for Colon Cancer Mortality After Event of Recurrence \nby Age Categories",
           xlab = "Time (until death) \n in Days")
```

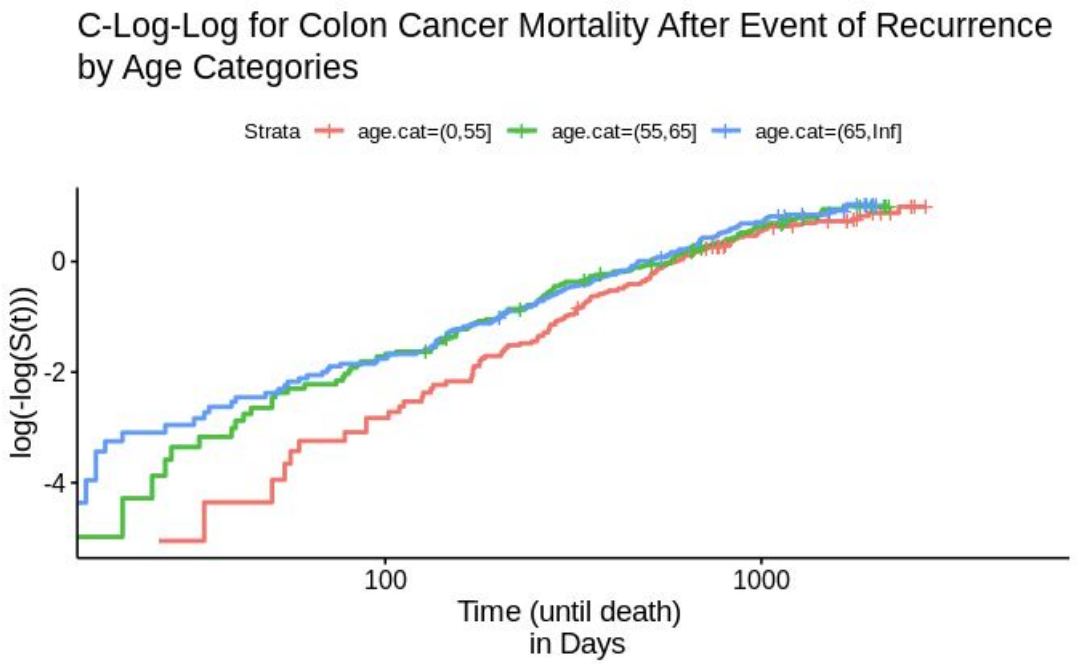

Figure 13: C-log-log plot for covariate Recurrece in the mortality model

In figure 13, the curves in the C-log-log plot are wider apart at the start of the study however they begin to narrow and overlap towards the end of the study. This pattern on the C-log-log causes great concern since it indicates that the cox proportional hazards assumption is violated.

Lastly, we plotted the C-log-log plot for the covariate.

```
grx.fit <- survfit(Surv(time - tstart, status) ~ rx, data = colon.gap2)

ggsurvplot(grx.fit , conf.int = F,
           fun = "cloglog",
           xlim = c(20, 5000),
           title = "C-Log-Log for Colon Cancer Mortality After Event of Recurrence \nby Treatment",
           xlab = "Time (until death) \n in Days")
```

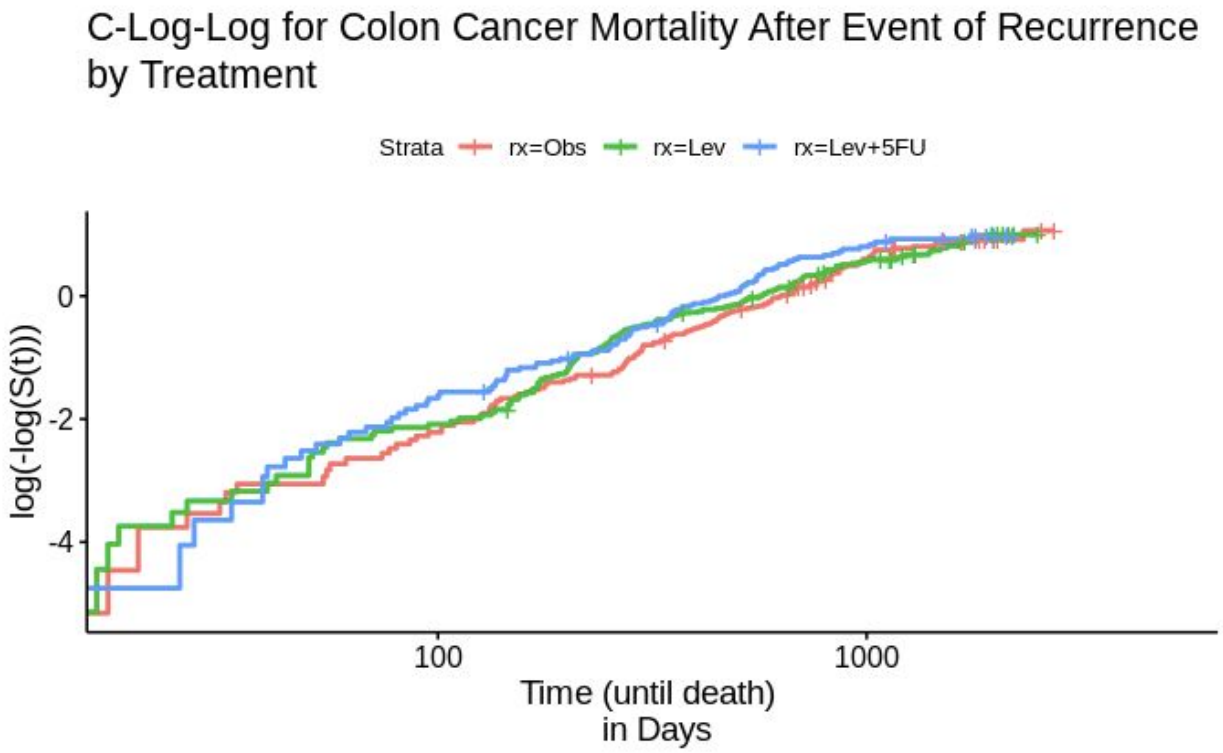

Figure 14: C-log-log plot for covariate Rx in the mortality model

From Figure 14, we conclude that the covariate rx does not violate the cox proportional hazards assumption since the curves are each roughly parallel to each other and follow the same trend.

Goodness of Fit Test

We then used the cox.zph function to further justify our conclusion.

```
gap.fit <- coxph(Surv(time - tstart, status) ~ node4 + age + rx , data=colon.gap2)
gap.zp <- cox.zph(gap.fit)
gap.zp
##         chisq df           p
## node4  0.391  1 0.532
## age   5.550  1 0.018
## rx     3.505  2 0.173
## GLOBAL 9.844  4 0.043
```

Since the p value for age is less than 0.05, there is significant evidence that the cox proportional model assumption is violated which aligns with our previous conclusion from the C-log-log plot for age in Figure 12.

We suspect that age may be a time varying covariate due its significance in the Cox Goodness of Fit Test. In order to confirm age is a time-varying coefficient we plot the residuals for age that derive from the Cox Goodness of Fit Test.

Residual Test

```
# a plot for the second variable in the fit
plot(gap.zp[2])
abline(0,0, col=2)
abline(h = gap.fit$coef[2], col=3, lwd=2, lty=2)
```

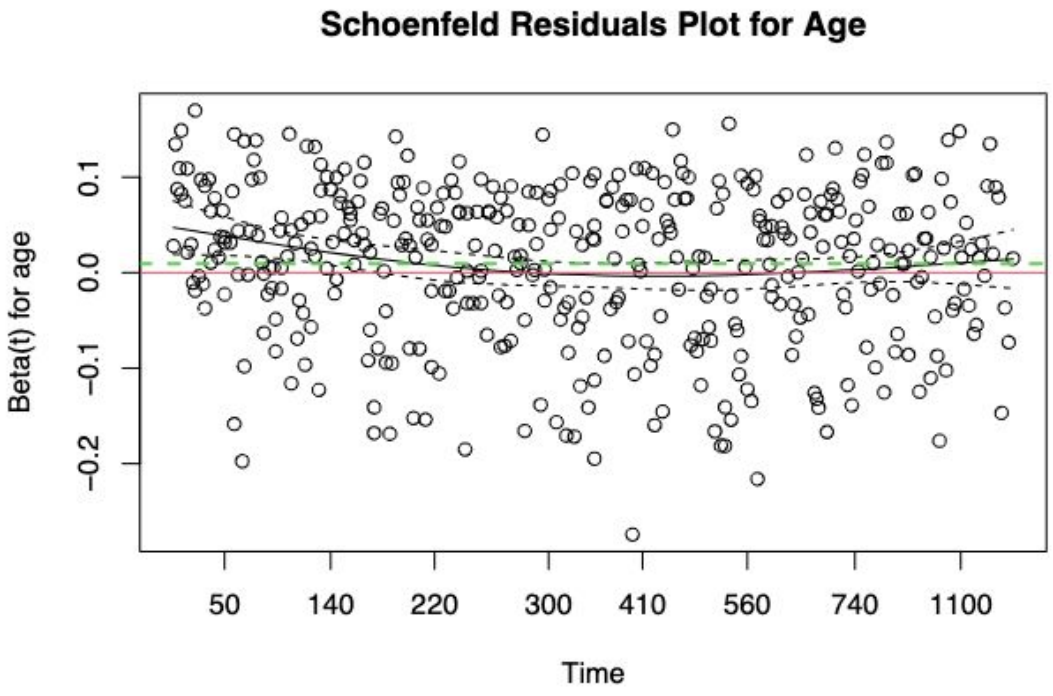

Figure 15: Schoenfeld Residuals Plot for covariate age in the gap model

Based on the in Figure 15, we conclude that age is not a time varying covariate during the study since there are no significant turning points in the Schoenfeld Residuals curve.

The Final Model

We concluded that it is best to stratify on the age categories which will estimate separate baselines for each of the three age groups. In this stratified model, age is no longer a parameter in the MLE thus we can no longer check whether age is a significant covariate as it was originally determined in the forward selection process.

Hence, the final model for the survival probability after the event of a recurrence is:

$$\text{Surv(time, status)} \sim \text{node4} + \text{strata(age.cat)} + \text{rx}$$

```
g.final.model <- coxph(Surv(time- tstart, status) ~ node4 + strata(age.cat) + rx, data=colon.gap2)

summary(g.final.model)

## Call:
## coxph(formula = Surv(time - tstart, status) ~ node4 + strata(age.cat) +
##     rx, data = colon.gap2)
##
##   n= 461, number of events= 409
##
##                 coef exp(coef) se(coef)  z Pr(>|z|)
## node41       0.51582   1.67501  0.10243 5.036 4.76e-07 ***
## rxLev        0.08346   1.08705  0.11548 0.723   0.4698
## rxLev+5FU 0.29178   1.33880  0.12817 2.276   0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## node41         1.675    0.5970  1.3703  2.047
## rxLev          1.087    0.9199  0.8669  1.363
## rxLev+5FU   1.339    0.7469  1.0414  1.721
##
## Concordance= 0.588  (se = 0.015 )
## Likelihood ratio test= 30.53  on 3 df,   p=1e-06
## Wald test                = 31.64  on 3 df,   p=6e-07
## Score (logrank) test = 32.28  on 3 df,   p=5e-07
```

From the results of the summary function for the final gap model, we see that after we control for all other significant covariates in the model the coefficient for covariate rxLev is 0.08346 with a p-value of 0.4698. The hazard ratio between the group treated with rxLev and the observation group is 1.08705. Additionally, the confidence interval for the hazard ratio is (0.8669, 1.363).

On the other hand, the coefficient for covariate rxLev is 0.29178 with a p-value of 0.0228. Moreover, the hazard ratio between the group treated with rexLev+5Fu and the observation group is 1.33880 and the confidence interval for the hazard ratio is (1.0414, 1.721).

Research Question 3

As an extension to the recurrence marginal model, we observed a gap model for the survival probability of patients after the event of recurrence. The final gap model allows us to determine whether the treatments Levamisole and Levamisole+5-FU still improve the survival rate of colon cancer patients after a recurrence.

Notably, our result indicates that there is no significant evidence to indicate a difference between the hazard rate of the group treated with Levamisole and the hazard rate of the observation group since the p-value is 0.4698 which is greater than 0.05. Additionally, the hazard rate for the group treated with Levamisole is only 8% greater than the hazard rate of the observation group. Which further justifies that the survival probability is similar for patients treated with Levamisole and observed patients with no treatment. Therefore, we once again conclude that the treatment Levamisole is not effective on improving the survival probability in colon cancer patients even after the event of recurrence.

Moreover, our results in the gap model illustrate that the hazard rate for the group treated with Levamisole+5-FU is roughly 34% greater than the hazard rate for the observation group. Since the p-value for the coefficient is 0.0228 which is less than 0.05, this means that there is a significant difference between the hazard rate of the group treated with Levamisole+5-FU and the hazard rate for the observation group. However, it is important to note that the coefficient for Levamisole+5-FU is greater than 0 which means that the hazard rate is greater than the hazard rate for the observed group.

Therefore, we conclude that the treatment Levamisole+5-FU actually hinders the survival probability for colon cancer patients with a recurrence. Hence, not only is the treatment Levamisole+5-FU no longer effective in improving a patient's survival rate but it can negatively affect them after a recurrence of cancer.

Conclusion

Our survival analysis has shown that the covariates node4, nodes, extent and surg are all significant in both marginal models for mortality and recurrence. We concluded that the treatment Levamisole is not effective on improving survival probability nor effective on delaying recurrence, whereas the treatment Levamisole+5-FU is both effective on improving survival probability and delaying recurrence for colon cancer patients. Yet, while it appears that Levamisole+5-FU is effective throughout the study, we determined that the treatment is no longer effective once a patient has a recurrent event of colon cancer. From the summary of the gap model, we inferred that Levamisole+5-FU may potentially have a negative effect in which patients treated with Levamisole+5-FU have a 30% higher hazard rate than observed patients with no treatment. It appears that Levamisole+5-FU only helps delay recurrence which may have influenced the results from the marginal model for mortality. In other words, since the treatment Levamisole+5-FU helps delay the event of recurrence then colon cancer patients can potentially

live longer. It is also worth mentioning that Levamisole alone did not have a significant effect on the survival probability after the event of recurrence.

Reference

JA Laurie, CG Moertel, TR Fleming, HS Wieand, JE Leigh, J Rubin, GW McCormack, JB Gerstner, JE Krook and J Malliard. Surgical adjuvant therapy of large-bowel carcinoma: An evaluation of levamisole and the combination of levamisole and fluorouracil: The North Central Cancer Treatment Group and the Mayo Clinic. J Clinical Oncology, 7:1447-1456, 1989.