Survival Analysis of Mortality of Adjuvant Chemotherapy for Colon Cancer (AIC)

Yijia Jiang, Ziyan Xu

2022-11-28

1. Data import and Examination

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.

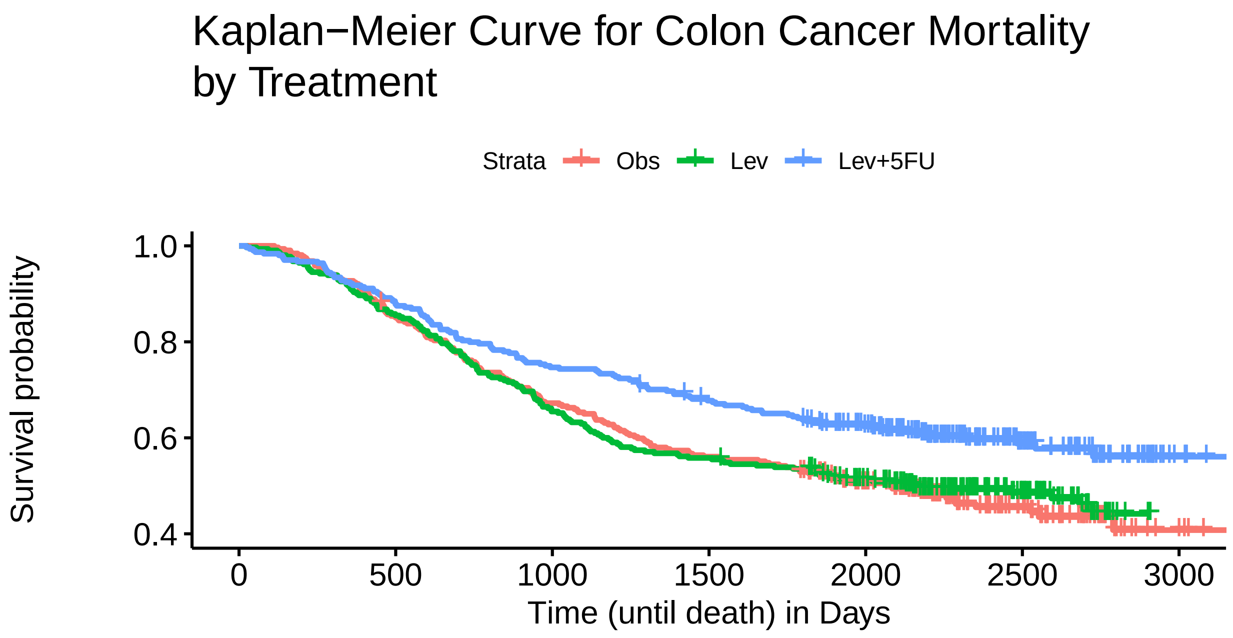
```
# import the dataset from survival package
data(cancer, package = "survival")
colon <- as_tibble(colon)

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

2. Kaplan-Meier Survival Estimate

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

ggsurvplot(death.fit, conf.int = F, break.time.by = 500, ylim = c(0.4,1.0),
           font.x.size = 12, font.y.size = 12, font.legend.size = 9, legend.labs = c("Obs","Lev","Lev+5F
           title = "Kaplan-Meier Curve for Colon Cancer Mortality \nby Treatment",
           xlab = "Time (until death) in Days",
           risk.table = T, risk.table.height = 0.25, risk.table.fontsize = 4,
           tables.theme = theme_cleantable())
```


Kaplan−Meier Curve for Colon Cancer Mortality by Treatment

Number at risk

Obs	315	267	211	176	141	50	6
Lev	310	265	203	173	145	58	4
Lev+5FU	304	267	227	203	170	65	7

From the plot above, 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.

3. Log-Rank Test

We do a proper Log-rank hypothesis test to test the null hypothesis of no difference among the three treatments in the mortality model.

```
d.rx.coxph <- coxph(Surv(time, status) ~rx, data = colon.death)
summary(d.rx.coxph)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = colon.death)
##
##   n= 929, number of events= 452
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev       -0.02664   0.97371  0.11030 -0.241  0.80917
## rxLev+5FU -0.37171   0.68955  0.11875 -3.130  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9737      1.027    0.7844    1.2087
## rxLev+5FU    0.6896      1.450    0.5464    0.8703
##
## Concordance= 0.536  (se = 0.013 )
## Likelihood ratio test= 12.15  on 2 df,   p=0.002
## Wald test            = 11.56  on 2 df,   p=0.003
## Score (logrank) test = 11.68  on 2 df,   p=0.003
```

From this log-rank test, we get a p-value that is closed to 0.002, which is significant at a 0.05 level. We want to conclude that there is a significant difference among the three treatments in the mortality model.

4. Tidy the Data for Model Development

Moreover, we notice that variables nodes and node4 both indicate similar information regarding the amount of positive lymph nodes an individual has. The variable nodes measures the number of lymph nodes with detectable cancer while node4 indicates whether there are more than 4 positive lymph nodes (0 = No, 1 =Yes). Therefore, we decided to use only the variable node4 in our analysis.

Additionally, 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 4.41% 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.

```
# check for missing values
apply(is.na(colon.death), 2, which) %>%
  str()
```

```
## List of 16
##  $ id      : int(0)
##  $ study   : int(0)
##  $ rx      : int(0)
##  $ sex     : int(0)
##  $ age     : int(0)
##  $ obstruct: int(0)
##  $ perfor  : int(0)
##  $ adhere  : int(0)
##  $ nodes   : int [1:18] 94 99 143 189 199 338 358 365 383 502 ...
##  $ status  : int(0)
##  $ differ  : int [1:23] 64 83 90 161 190 200 202 244 294 325 ...
##  $ extent  : int(0)
##  $ surg    : int(0)
##  $ node4   : int(0)
##  $ time    : int(0)
##  $ etype   : int(0)
```

```
# remove missing values
colon.death1 <- na.omit(colon.death)
```

5. Cox PH Model

5.1 Model 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 AIC and BIC value.

```
# fit all the variables in the model
d.model.full <- coxph(Surv(time, status) ~ sex + age + obstruct + perfor + adhere + differ + extent + su

# stepwise selection with AIC criterion
d.model.aic <- step(d.model.full, direction = "both", k = 2)
```

```
## Start:  AIC=5425.36
## Surv(time, status) ~ sex + age + obstruct + perfor + adhere +
##     differ + extent + surg + node4 + rx
##
##            Df    AIC
## - perfor    1 5423.4
## - sex       1 5423.4
## - adhere    1 5425.0
## <none>        5425.4
## - differ    1 5425.9
## - age       1 5426.7
## - surg      1 5427.8
## - obstruct  1 5428.0
## - rx        2 5431.9
## - extent    1 5439.6
## - node4     1 5502.9
##
## Step:  AIC=5423.37
## Surv(time, status) ~ sex + age + obstruct + adhere + differ +
##     extent + surg + node4 + rx
##
##            Df    AIC
## - sex       1 5421.4
## - adhere    1 5423.1
## <none>        5423.4
## - differ    1 5423.9
## - age       1 5424.7
## + perfor    1 5425.4
## - surg      1 5425.8
## - obstruct  1 5426.1
## - rx        2 5429.9
## - extent    1 5437.7
## - node4     1 5501.0
##
## Step:  AIC=5421.42
## Surv(time, status) ~ age + obstruct + adhere + differ + extent +
##     surg + node4 + rx
##
##            Df    AIC
## - adhere    1 5421.1
## <none>        5421.4
## - differ    1 5422.0
## - age       1 5422.8
## + sex       1 5423.4
## + perfor    1 5423.4
## - surg      1 5423.9
## - obstruct  1 5424.2
## - rx        2 5428.0
## - extent    1 5435.7
## - node4     1 5499.1
##
## Step:  AIC=5421.13
## Surv(time, status) ~ age + obstruct + differ + extent + surg +
##     node4 + rx
##
##            Df    AIC
## <none>        5421.1
## + adhere    1 5421.4
## - differ    1 5422.2
## - age       1 5423.0
## + perfor    1 5423.1
## + sex       1 5423.1
## - surg      1 5423.7
## - obstruct  1 5423.9
## - rx        2 5427.9
## - extent    1 5436.4
## - node4     1 5498.3
```

The resulting model with the lowest AIC is: Surv(time, status) ~ age + obstruct + differ + extent + surg + node4 + rx

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.model.aic)
```

```
## 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
## age      -2767.7  0.4782  1   0.489232
## obstruct -2765.4  4.5874  1   0.032207 *
## differ   -2760.8  9.2216  1   0.002392 **
## extent   -2749.4 22.6937  1    1.9e-06 ***
## surg     -2747.6  3.6817  1   0.055014 .
## node4    -2707.9 79.2931  1  < 2.2e-16 ***
## rx       -2702.6 10.7336  2   0.004669 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that p-values for covariates “obstruct, differ, extent, surg, node4 and rx” are much smaller than 0.05, indicating that they have a significant effect on time until death. Therefore, we will include these 6 covariates in our Cox PH model.

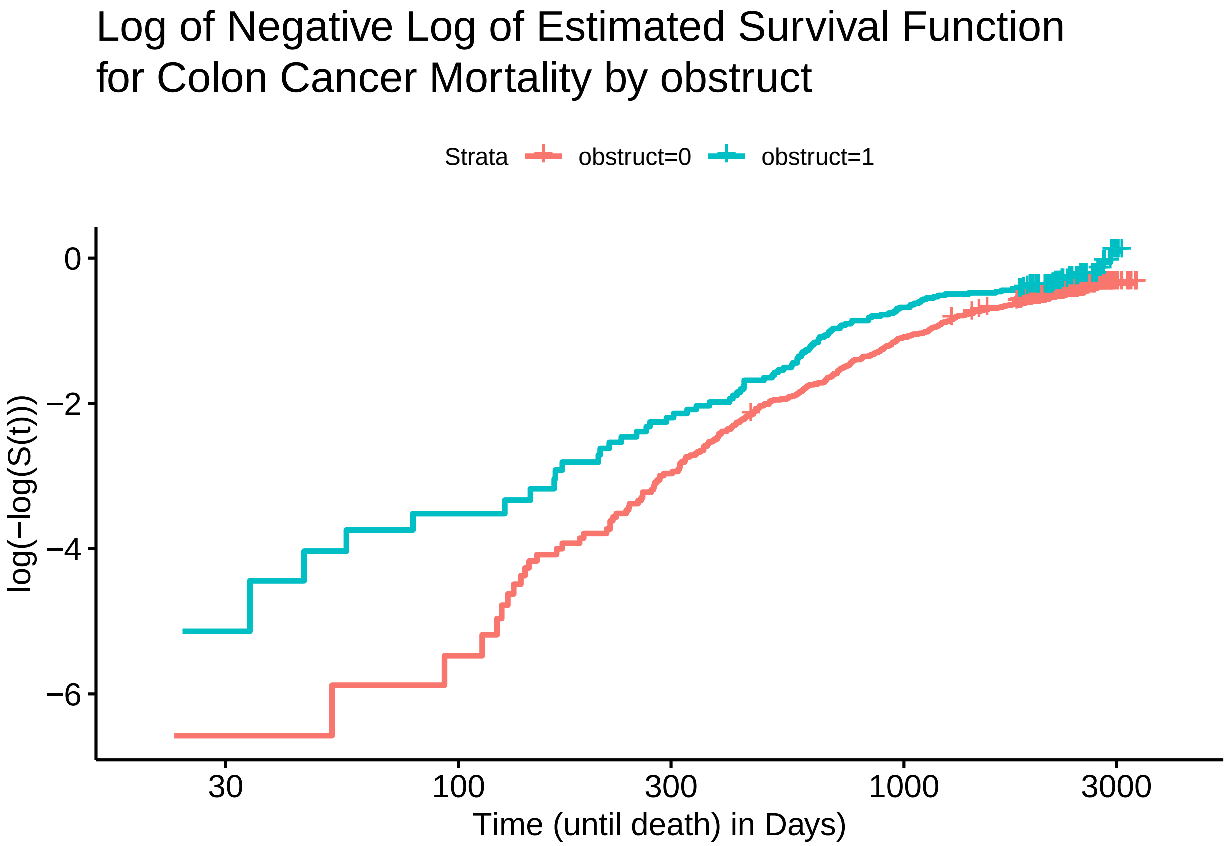
Therefore, the resulting model is: Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx

5.2 Model Diagnostic

5.2.1 Log of Negative Log of Estimated Survival Function

To check the proportional hazards assumption for this model, we use a diagnostic plot such as the log of negative log of estimated survival function as we did for covariate treatment above. Then we checked the significance of each covariate again using Analysis of Deviance procedure to ensure that our previous model is still valid. First comes to covariate obstruct.

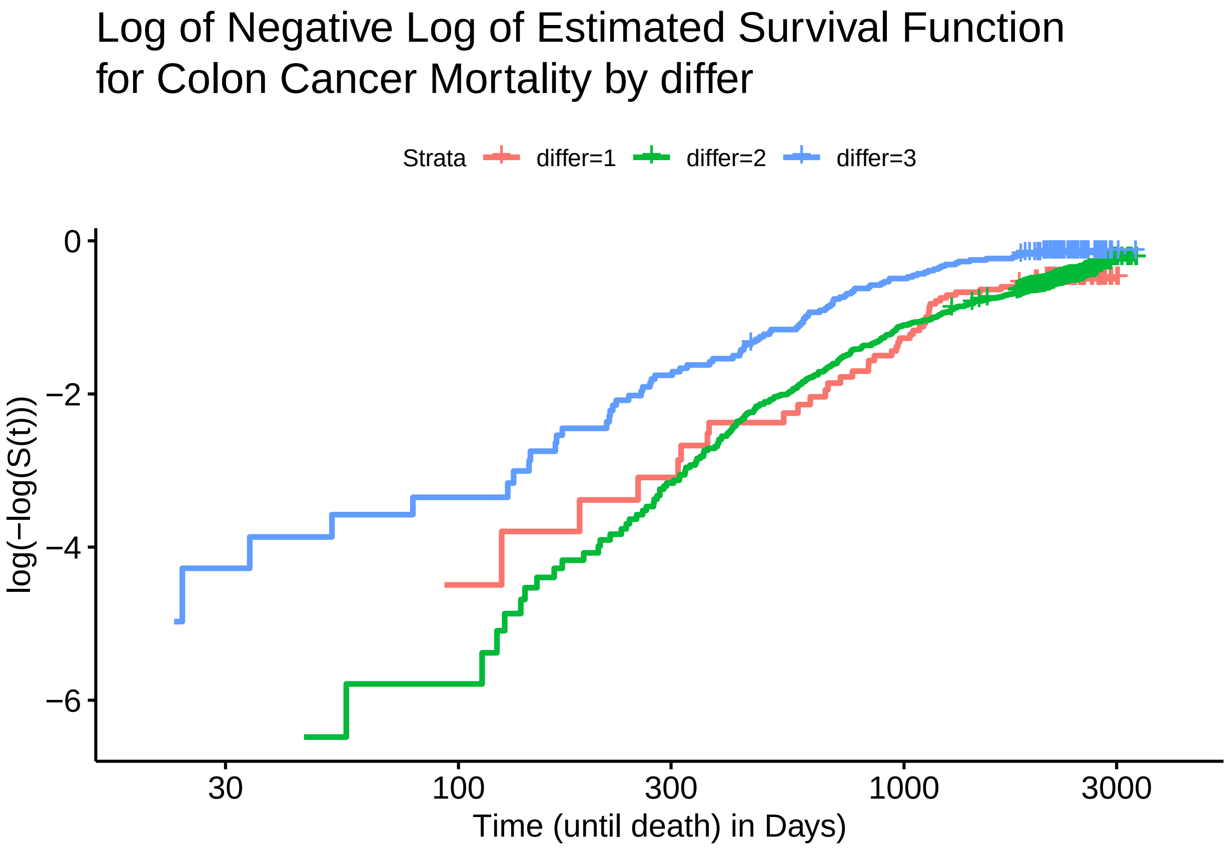
```
d.obstruct.fit <- survfit(Surv(time, status) ~ obstruct, data = colon.death1)
ggsurvplot(d.obstruct.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by
```



According to the plot, the distance between two obstruct curves begin to narrow after 2000 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 2000 days. Hence we believe it is best to ignore the noisiness of the plot since the curves are roughly parallel after 2000 days. Thus, the cox proportional hazards assumption is valid to use for this covariate.

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

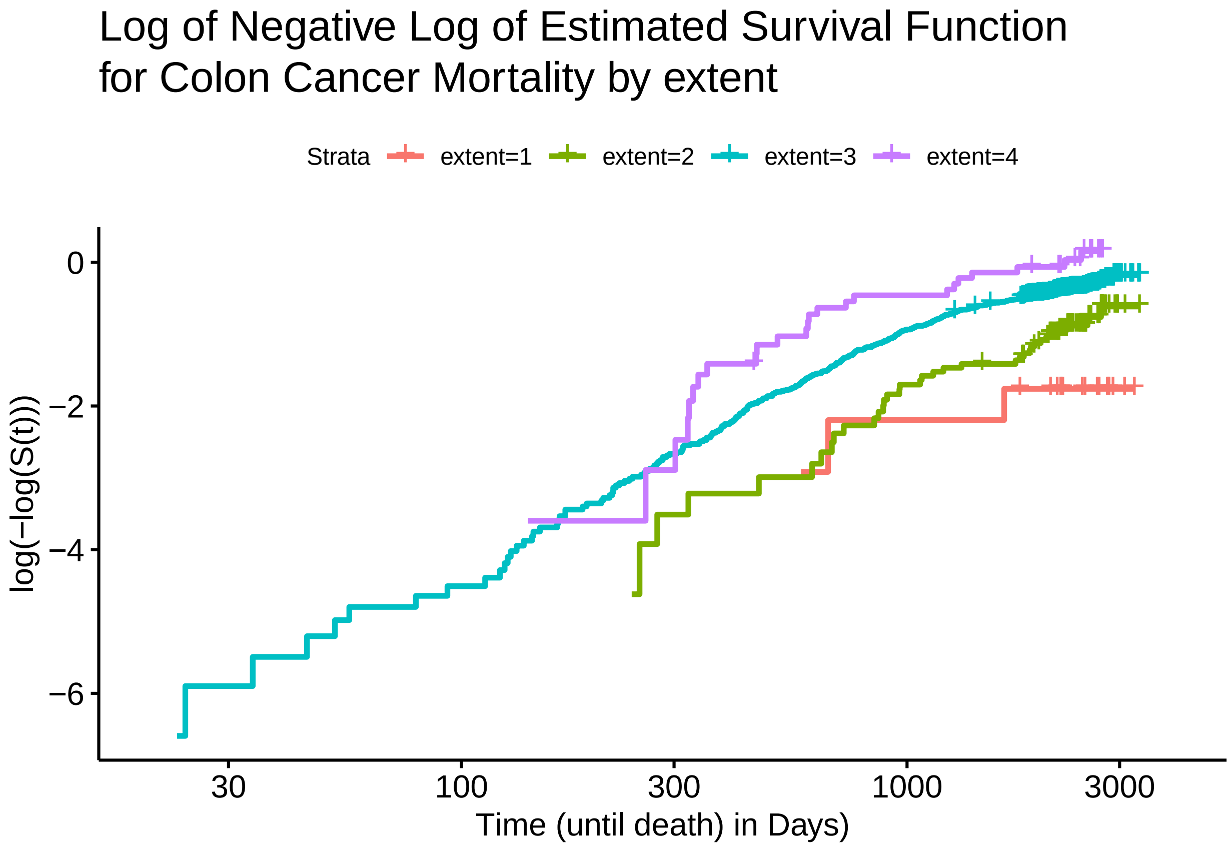
```
d.differ.fit <- survfit(Surv(time, status) ~ differ, data = colon.death1)
ggsurvplot(d.differ.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by d
```



According to the plot, the curves in the C-log-log plot are crossing over after 300 days. Since there are not enough data points in each differ group to show a more comprehensive trend, it's hard for us to make a decision based on the plot.

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

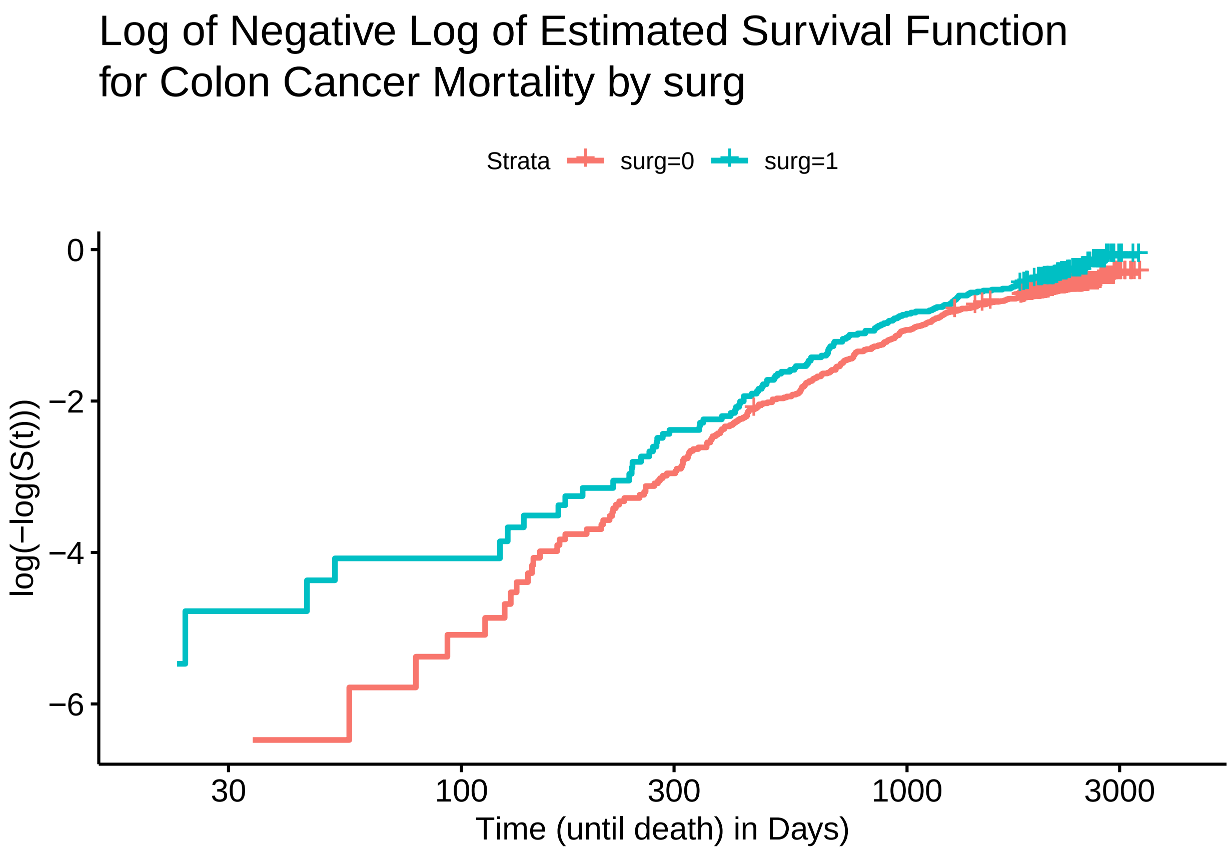
```
d.extent.fit <- survfit(Surv(time, status) ~ extent, data = colon.death1)
ggsurvplot(d.extent.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by e
```



According to the plot, 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.

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

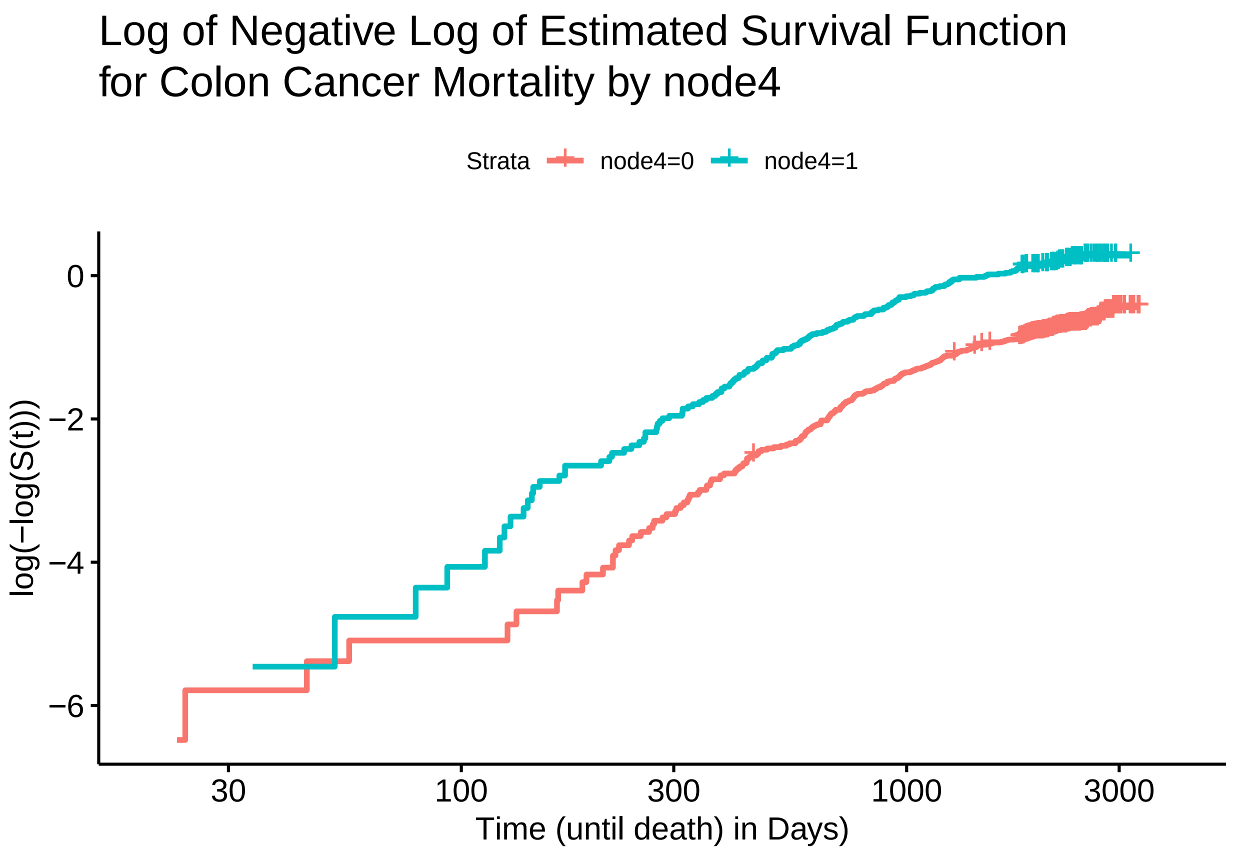
```
d.surg.fit <- survfit(Surv(time, status) ~ surg, data = colon.death1)
ggsurvplot(d.surg.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by
```



According to the plot, the distance between two surg curves begin to narrow after 2000 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 2000 days. Hence we believe it is best to ignore the noisiness of the plot since the curves are roughly parallel after 2000 days. Thus, the cox proportional hazards assumption is valid to use for this covariate.

We continue to plot the C-log-log plot for the covariate node4.

```
d.node4.fit <- survfit(Surv(time, status) ~ node4, data = colon.death1)
ggsurvplot(d.node4.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size = 9,
           fun = "cloglog",
           xlim = c(20, 4000),
           xlab = "Time (until death) in Days)",
           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer Mortality by
```



According to the plot, 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.

5.2.2 Test Interaction for Proportionality

```
d.model.aic.inter <- coxph(Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx + obstruc
summary(d.model.aic.inter)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ obstruct + differ + extent +
##     surg + node4 + rx + obstruct * log(time) + differ * log(time) +
##     extent * log(time) + surg * log(time) + node4 * log(time) +
##     rx * log(time), data = colon.death1)
##
##   n= 888, number of events= 430
##
##                            coef  exp(coef)   se(coef)       z Pr(>|z|)
## obstruct             -8.898e-01  4.108e-01  1.599e-01  -5.566 2.61e-08 ***
## differ               -1.887e-01  8.280e-01  1.206e-01  -1.565 0.117609
## extent               -5.128e-01  5.988e-01  1.505e-01  -3.407 0.000658 ***
## surg                 -1.587e+00  2.046e-01  1.394e-01 -11.384  < 2e-16 ***
## node4                -9.564e-01  3.843e-01  1.338e-01  -7.146 8.95e-13 ***
## rxLev                -6.266e-01  5.344e-01  1.334e-01  -4.695 2.66e-06 ***
## rxLev+5FU            1.098e+00  2.998e+00  1.448e-01   7.580 3.45e-14 ***
## log(time)           -1.744e+02  1.793e-76  9.465e+00 -18.427  < 2e-16 ***
## obstruct:log(time)   1.344e-01  1.144e+00  2.366e-02   5.679 1.35e-08 ***
## differ:log(time)     2.659e-02  1.027e+00  1.786e-02   1.489 0.136476
## extent:log(time)     7.306e-02  1.076e+00  2.184e-02   3.346 0.000821 ***
## surg:log(time)       2.510e-01  1.285e+00  2.039e-02  12.311  < 2e-16 ***
## node4:log(time)      1.530e-01  1.165e+00  1.988e-02   7.695 1.41e-14 ***
## rxLev:log(time)      7.353e-02  1.076e+00  1.953e-02   3.765 0.000166 ***
## rxLev+5FU:log(time) -1.737e-01  8.406e-01  2.102e-02  -8.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     exp(coef) exp(-coef) lower .95 upper .95
## obstruct            4.108e-01  2.435e+00 3.003e-01 5.619e-01
## differ              8.280e-01  1.208e+00 6.537e-01 1.049e+00
## extent              5.988e-01  1.670e+00 4.459e-01 8.043e-01
## surg                2.046e-01  4.888e+00 1.557e-01 2.688e-01
## node4               3.843e-01  2.602e+00 2.956e-01 4.996e-01
## rxLev               5.344e-01  1.871e+00 4.114e-01 6.942e-01
## rxLev+5FU           2.998e+00  3.336e-01 2.257e+00 3.982e+00
## log(time)           1.793e-76  5.576e+75 1.574e-84 2.043e-68
## obstruct:log(time)  1.144e+00  8.743e-01 1.092e+00 1.198e+00
## differ:log(time)    1.027e+00  9.738e-01 9.916e-01 1.064e+00
## extent:log(time)    1.076e+00  9.295e-01 1.031e+00 1.123e+00
## surg:log(time)      1.285e+00  7.780e-01 1.235e+00 1.338e+00
## node4:log(time)     1.165e+00  8.581e-01 1.121e+00 1.212e+00
## rxLev:log(time)     1.076e+00  9.291e-01 1.036e+00 1.118e+00
## rxLev+5FU:log(time) 8.406e-01  1.190e+00 8.066e-01 8.759e-01
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 4901   on 15 df,   p=<2e-16
## Wald test            = 983.6  on 15 df,   p=<2e-16
## Score (logrank) test = 2531   on 15 df,   p=<2e-16
```

Based on the test interaction for proportionality, the result shows that all the interactions of covaraites with `log(time)` are significant (i.e., less than 0.05). Therefore, we can conclude that the proportionality assumptions for all the covariates are met.

5.3 The final model

With the inclusion of treatment, our final model is given by: Surv(time, status) ~ differ + extent + surg + node4 + rx.

```
d.final.model <-coxph(Surv(time, status) ~ differ + extent + surg + node4 + rx, data = colon.death1)
summary(d.final.model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ differ + extent + surg +
##     node4 + rx, data = colon.death1)
##
##   n= 888, number of events= 430
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## differ     0.17411   1.19019  0.09962  1.748  0.08052 .
## extent     0.49535   1.64108  0.11827  4.188 2.81e-05 ***
## surg       0.23813   1.26887  0.10588  2.249  0.02451 *
## node4      0.91032   2.48512  0.10044  9.063  < 2e-16 ***
## rxLev     -0.03101   0.96947  0.11383 -0.272  0.78532
## rxLev+5FU -0.36002   0.69766  0.12162 -2.960  0.00307 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## differ       1.1902     0.8402    0.9791    1.4468
## extent       1.6411     0.6094    1.3015    2.0692
## surg         1.2689     0.7881    1.0311    1.5615
## node4        2.4851     0.4024    2.0410    3.0258
## rxLev        0.9695     1.0315    0.7756    1.2118
## rxLev+5FU    0.6977     1.4334    0.5497    0.8855
##
## Concordance= 0.657  (se = 0.013 )
## Likelihood ratio test= 123  on 6 df,   p=<2e-16
## Wald test            = 126.3  on 6 df,   p=<2e-16
## Score (logrank) test = 133.9  on 6 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 -0.031 with a p-value of 0.78, and the hazard ratio between the group treated with rxLev and the observation group is 0.97. Furthermore, the confidence interval for the hazard ratio is (0.78, 1.21). On the other hand, the coefficient for covariate rxLev is -0.36 with a p-value of 0.003, and the hazard ratio between the group treated with rexLev+5Fu and the observation group is 0.678. Furthermore, the confidence interval for the hazard ratio is (0.55, 0.89).