Survival Analysis of Mortality of Adjuvant Chemotherapy for Colon Cancer

Yijia Jiang, Ziyan Xu

2022-11-30

1. Data Tidy

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.

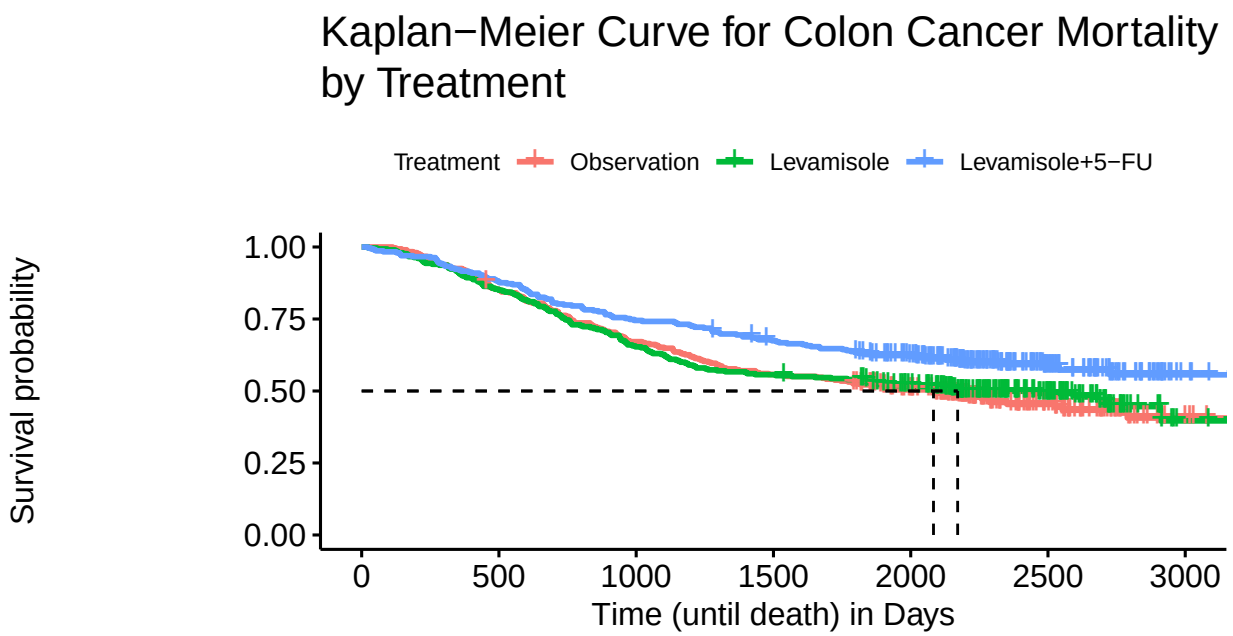
```
# import the dataset from survival package
data(cancer, package = "survival")
colon <- as_tibble(colon)

# tidy mortality dataset
colon.death <- colon %>%
  dplyr::select(-id, -study, -nodes) %>%
  drop_na() %>%
  mutate(rx = as.numeric(rx)) %>%
  subset(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,
           font.x.size = 12, font.y.size = 12, font.legend.size = 9, surv.median.line = "hv",
           legend.title = "Treatment", legend.labs = c("Observation", "Levamisole","Levamisole+5-FU"),
           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())
```



Number at risk

Observation	308	261	206	171	139	49	6
Levamisole	300	255	196	167	142	57	4
Levamisole+5−FU	298	262	222	198	165	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. The median survival time for observation group and levamisole group are approximately 2100 days and 2200 days. However, until the end of the trial, the survival probability of Levamisole+5-FU treatment group is greater than 50% as we fail to observe the curve doesn't cross 50%, which means the median survival is simply undefined.

3. Log-Rank Test

Noticing the difference of survival probability between the three treatment groups, we do a Log-rank hypothesis test to test the null hypothesis of no difference among the three treatments in the mortality model.

```
survdiff(Surv(time, status) ~ rx, data = colon.death)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ rx, data = colon.death)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=1 308      165      145      2.67      3.99
## rx=2 300      154      141      1.13      1.66
## rx=3 298      122      154      6.77     10.44
##
##  Chisq= 10.6  on 2 degrees of freedom, p= 0.005
```

From this log-rank test, we get a p-value that is closed to 0.005, 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. Cox PH Model

4.1 Model Selection

We now use automatic stepwise selection with Akaike information criterion (AIC) to determine the covariates that best represent an appropriate cox proportional hazards model for the event of death.

```
# fit all the variables in the model
d.model.full <- coxph(Surv(time, status) ~ . -etype, data = colon.death)

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

```
## Start:  AIC=5582.07
## Surv(time, status) ~ (rx + sex + age + obstruct + perfor + adhere +
##     differ + extent + surg + node4 + etype) - etype
##
##            Df    AIC
## - perfor    1 5580.1
## - sex       1 5580.2
## - adhere    1 5581.5
## <none>        5582.1
## - age       1 5583.8
## - differ    1 5583.8
## - surg      1 5585.1
## - obstruct  1 5585.4
## - rx        1 5588.3
## - extent    1 5594.0
## - node4     1 5661.2
##
## Step:  AIC=5580.09
## Surv(time, status) ~ rx + sex + age + obstruct + adhere + differ +
##     extent + surg + node4
##
##            Df    AIC
## - sex       1 5578.2
## - adhere    1 5579.6
## <none>        5580.1
## - age       1 5581.8
## - differ    1 5581.9
## + perfor    1 5582.1
## - surg      1 5583.1
## - obstruct  1 5583.6
## - rx        1 5586.3
## - extent    1 5592.2
## - node4     1 5659.3
##
```

```
## Step:  AIC=5578.21
## Surv(time, status) ~ rx + age + obstruct + adhere + differ +
##     extent + surg + node4
##
##            Df    AIC
## - adhere    1 5577.7
## <none>        5578.2
## - age       1 5579.9
## - differ    1 5580.1
## + sex       1 5580.1
## + perfor    1 5580.2
## - surg      1 5581.3
## - obstruct  1 5581.6
## - rx        1 5584.5
## - extent    1 5590.2
## - node4     1 5657.3
##
## Step:  AIC=5577.69
## Surv(time, status) ~ rx + age + obstruct + differ + extent +
##     surg + node4
##
##            Df    AIC
## <none>        5577.7
## + adhere    1 5578.2
## + perfor    1 5579.6
## + sex       1 5579.6
## - age       1 5579.7
## - differ    1 5580.1
## - surg      1 5581.0
## - obstruct  1 5581.1
## - rx        1 5584.0
## - extent    1 5590.9
## - node4     1 5656.5
```

The resulting model with the lowest AIC is: Surv(time, status) ~ obstruct + differ + extent + surg + node4 + age + 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 stepwise 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     -2847.5
## rx       -2843.0  9.0611  1  0.0026111 **
## age      -2842.7  0.5802  1  0.4462381
## obstruct -2840.3  4.7158  1  0.0298867 *
## differ   -2834.5 11.6373  1  0.0006464 ***
## extent   -2824.3 20.5506  1  5.807e-06 ***
```

```
## surg     -2822.3  3.9977  1  0.0455622 *
## node4    -2781.8 80.8270  1  < 2.2e-16 ***
## ---
## 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, except for "age", indicating that they have a significant effect on time until death. Therefore, we will include these 6 covariates in our Cox PH model.

Therefore, we obtain the resulting model, which is Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx.

4.2 Model Diagnostics

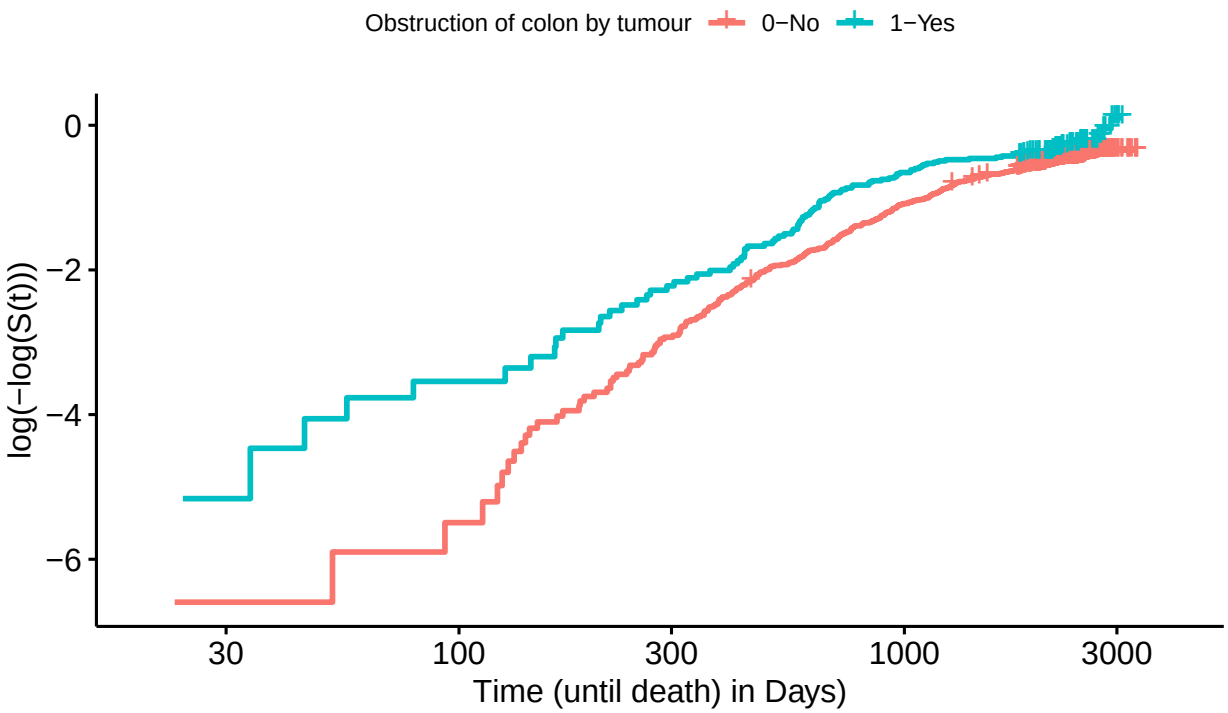
4.2.1 Check proportionality of hazard ratios

Log of Negative Log of Estimated Survival Function

To check the proportional hazards assumption for this model, we make diagnostic plots using log of negative log of estimated survival function. First comes to covariate obstruct.

```
d.obstruct.fit <- survfit(Surv(time, status) ~ obstruct, data = colon.death)
cloglog_obstruct <- ggsurvplot(d.obstruct.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.l
                               fun = "cloglog",
                               xlim = c(20, 4000),
                               xlab = "Time (until death) in Days)",
                               title = "Log of Negative Log of Estimated Survival Function \nfor Colon
                               legend.title = "Obstruction of colon by tumour",
                               legend.labs = c("0-No", "1-Yes"))

cloglog_obstruct
```

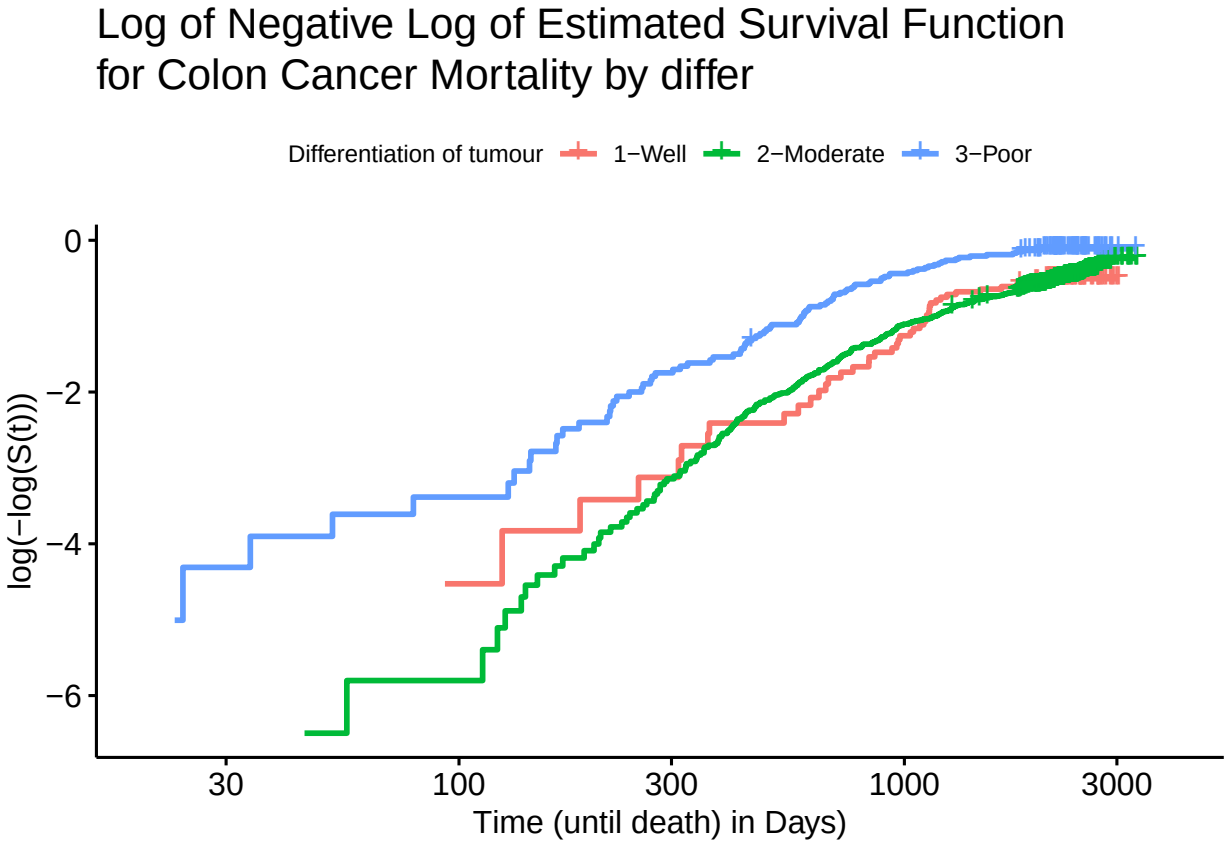


According to the plot, the distance between two obstruct curves begin to narrow after 2000 days which causes some concern that the assumption might be violated. However, it is also reasonable to assume that the curves are wider apparent earlier in the study since there are less occurrences of death before 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 "obstruct".

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

```
d.differ.fit <- survfit(Surv(time, status) ~ differ, data = colon.death)
cloglog_differ <- ggsurvplot(d.differ.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend
                             fun = "cloglog",
                             xlim = c(20, 4000),
                             xlab = "Time (until death) in Days)",
                             title = "Log of Negative Log of Estimated Survival Function \nfor Colon Ca
                             legend.title = "Differentiation of tumour",
                             legend.labs = c("1-Well", "2-Moderate","3-Poor"))

cloglog_differ
```

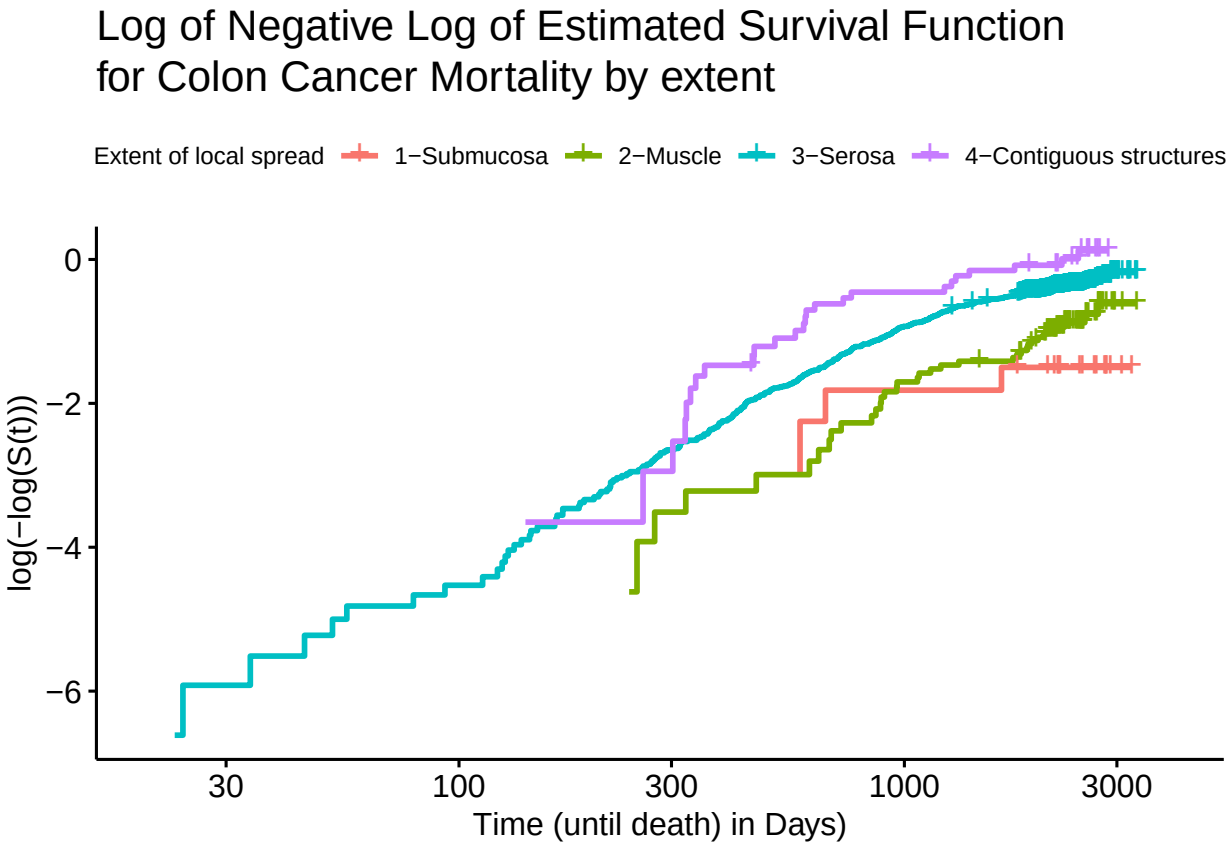


According to the plot, the curves in the C-log-log plot are crossing over after 300 days. The assumption of proportional hazard ratio between the three differ groups is violated as we fail to see three parallel lines against log time.

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

```
d.extent.fit <- survfit(Surv(time, status) ~ extent, data = colon.death)
cloglog_extent <- ggsurvplot(d.extent.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend
                             fun = "cloglog",
                             xlim = c(20, 4000),
                             xlab = "Time (until death) in Days)",
                             title = "Log of Negative Log of Estimated Survival Function \nfor Colon Can
                             legend.title = "Extent of local spread",
                             legend.labs = c("1-Submucosa", "2-Muscle", "3-Serosa", "4-Contiguous struct

cloglog_extent
```

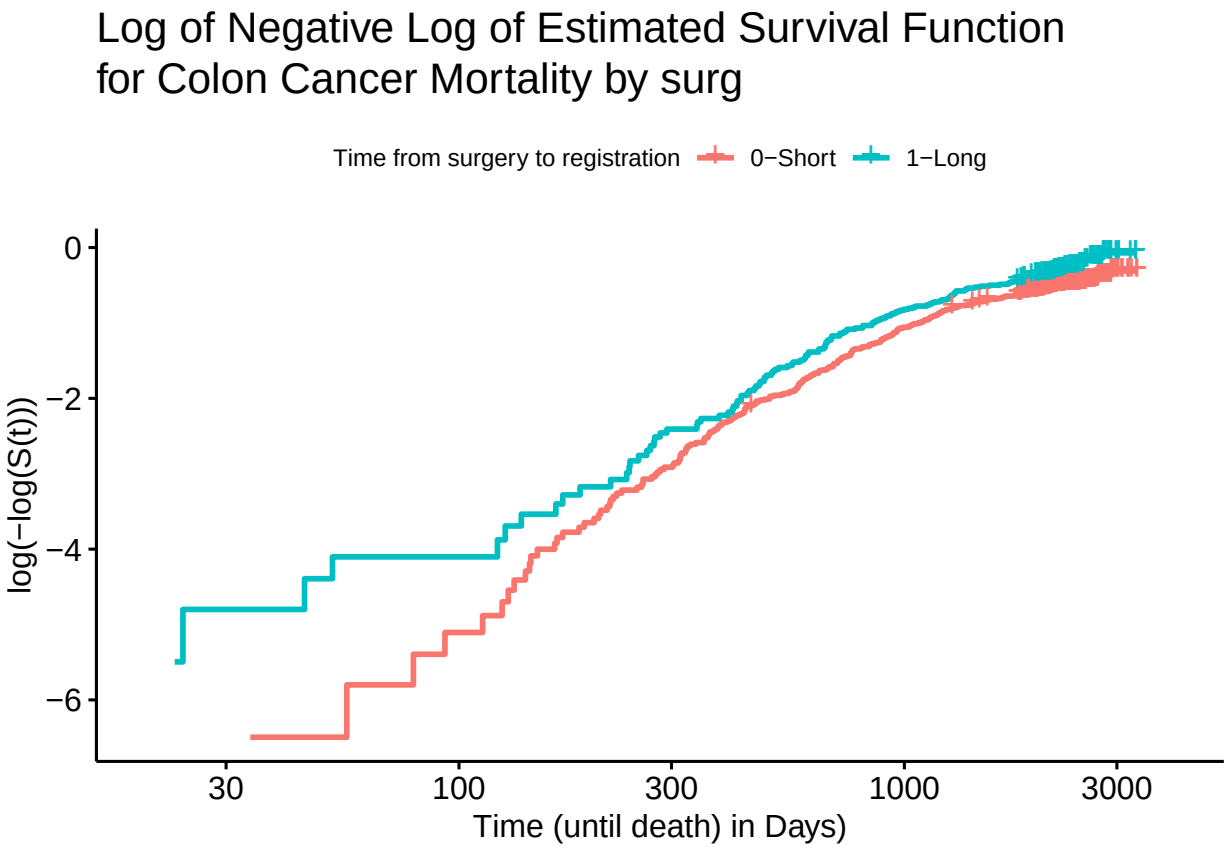


According to the plot, the curves in the C-log-log plot are crossing over after 100 days. The assumption of proportional hazard ratio of extent is violated as we fail to see four parallel lines against log time.

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

```
d.surg.fit <- survfit(Surv(time, status) ~ surg, data = colon.death)
cloglog_surg <- ggsurvplot(d.surg.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.si
                           fun = "cloglog",
                           xlim = c(20, 4000),
                           xlab = "Time (until death) in Days)",
                           title = "Log of Negative Log of Estimated Survival Function \nfor Colon Canc
                           legend.title = "Time from surgery to registration",
                           legend.labs = c("0-Short", "1-Long"))

cloglog_surg
```

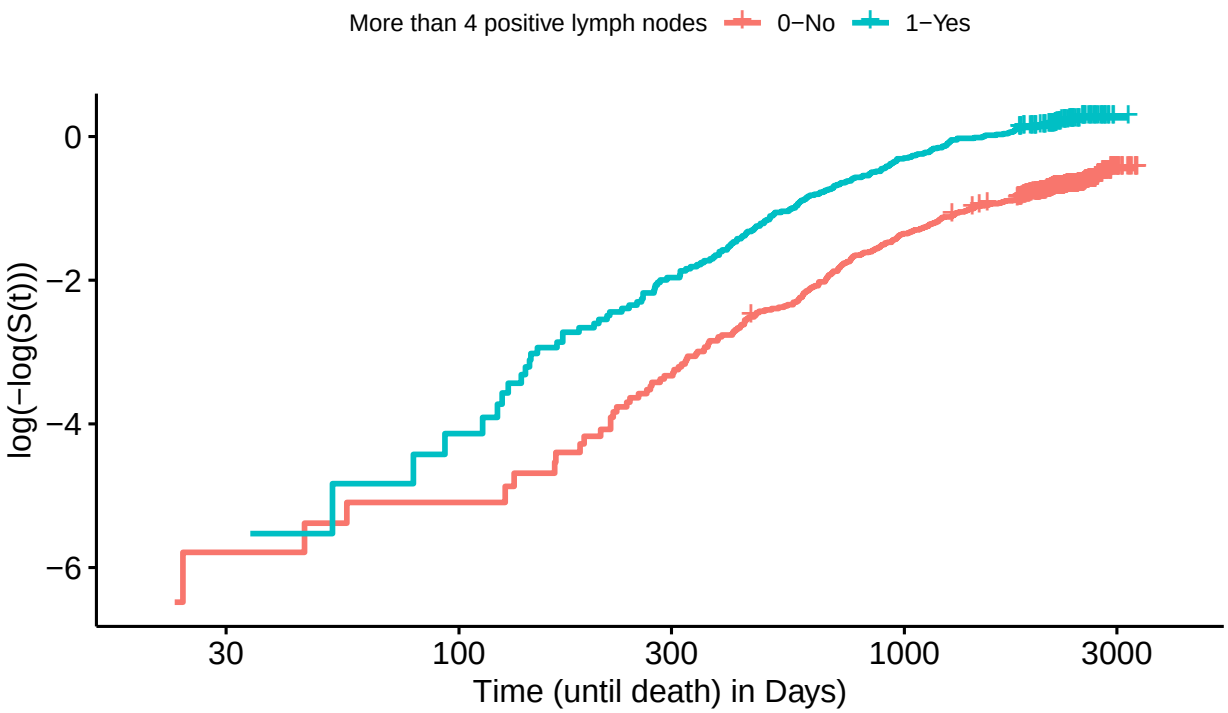


According to the plot, the distance between two surg curves begin to narrow after 100 days and becomes parallel to each other. We can assume that the cox proportional hazards assumption is valid to use for `surg`.

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

```
d.node4.fit <- survfit(Surv(time, status) ~ node4, data = colon.death)
cloglog_node4 <- ggsurvplot(d.node4.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.s
                            fun = "cloglog",
                            xlim = c(20, 4000),
                            xlab = "Time (until death) in Days)",
                            title = "Log of Negative Log of Estimated Survival Function \nfor Colon Canc
                            legend.title = "More than 4 positive lymph nodes",
                            legend.labs = c("0-No", "1-Yes"))

cloglog_node4
```

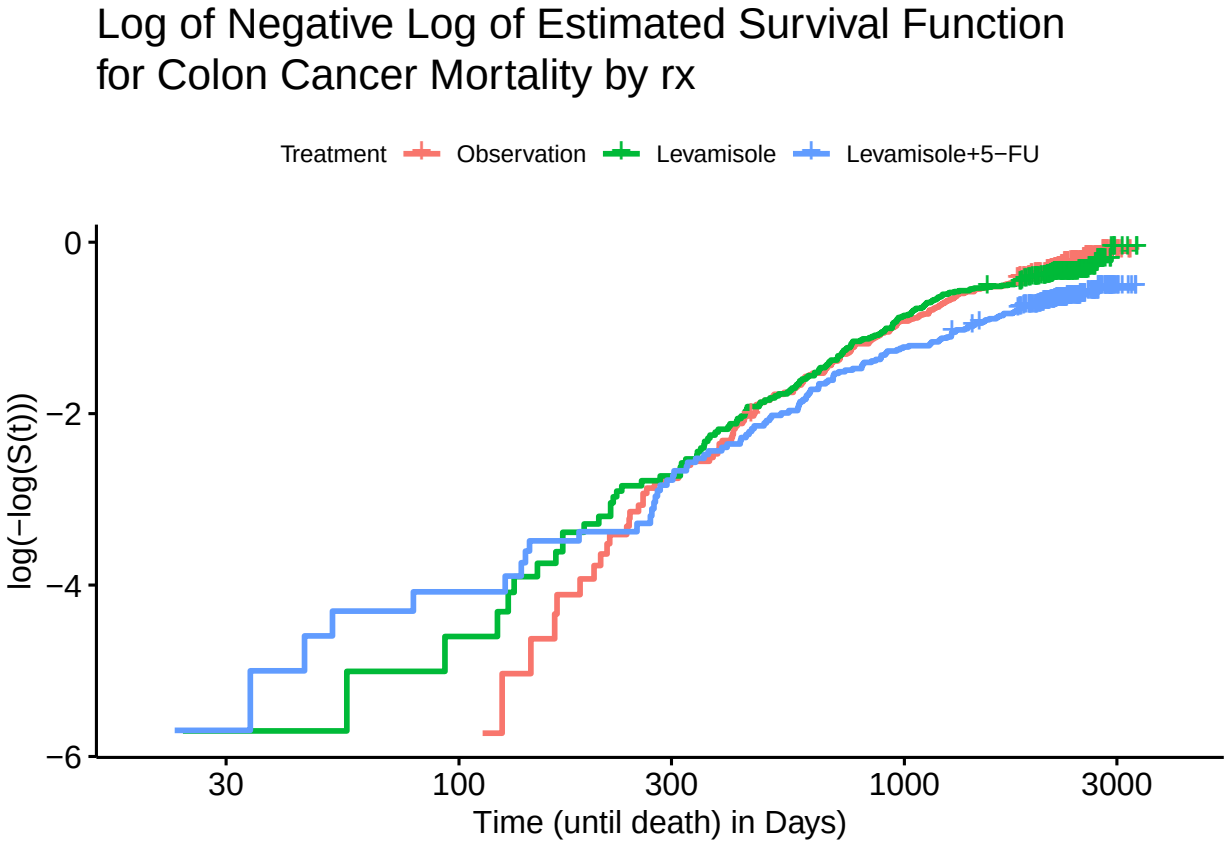


According to the plot, the two curves in this C-log-log plot cross over at the beginning of the study but appear to be parallel to each other after 100 days. Since the data is oftentimes noisy at the beginning of the study, the cross over does not cause too much concern. Overall, we believe that the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

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

```
d.surg.fit <- survfit(Surv(time, status) ~ rx, data = colon.death)
cloglog_rx <- ggsurvplot(d.surg.fit, conf.int = F, font.x.size = 12, font.y.size = 12, font.legend.size
                         fun = "cloglog",
                         xlim = c(20, 4000),
                         xlab = "Time (until death) in Days)",
                         title = "Log of Negative Log of Estimated Survival Function \nfor Colon Cancer
                         legend.title = "Treatment",
                         legend.labs = c("Observation", "Levamisole","Levamisole+5-FU"))

cloglog_rx
```



According to the plot, the distance between three treatment 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.

```
# combine all the cloglog plots and save them to a pdf file
splots <- list()
splots[[1]] <- cloglog_obstruct
splots[[2]] <- cloglog_differ
splots[[3]] <- cloglog_extent
splots[[4]] <- cloglog_surg
splots[[5]] <- cloglog_node4
splots[[6]] <- cloglog_rx
cloglog_plot = arrange_ggsurvplots(splots, print = FALSE, ncol = 2, nrow = 3)
ggsave(cloglog_plot,file = "./plot/d.C-log-log-plots.pdf",width = 12,height = 15)
ggsave(cloglog_plot,file = "./plot/d.C-log-log-plots.png",width = 12,height = 15)
```

Schoenfeld residuals

```
czph <- cox.zph(coxph(Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx, data = colon
schoenfeld_plot <- ggcoxzph(czph, font.main = 10, font.x = 10, font.y = 10, font.tickslab = 8,
                            point.alpha = 0.5, point.col = "grey25")
ggsave("./plot/d.schoenfeld residual_plots.pdf", arrangeGrob(grobs = schoenfeld_plot))
ggsave("./plot/d.schoenfeld residual_plots.png", arrangeGrob(grobs = schoenfeld_plot))
```

From the output above, the tests for covariates "obstruct", "node4", "differ" are statistically significant, and the global test is also statistically significant. Therefore, we can assume the violation of proportional hazards on these covariates, which requires corrections of non-proportional hazard ratio.

4.2.2 Test influential observations

```
# check outliers in term of dfbeta
outlier_dfbeta <- ggcoxdiagnostics(coxph(Surv(time, status) ~ obstruct + differ + extent + surg + node4
                 type = "dfbeta", linear.predictions = FALSE,ggtheme = theme_bw())

ggsave(outlier_dfbeta, file = "./plot/d.outlier_dfbeta.pdf")
ggsave(outlier_dfbeta, file = "./plot/d.outlier_dfbeta.png")
```

It's also possible to check outliers by visualizing the influence of each point, in terms of DFBETA - the impact on the coefficient of covariates in the model were that specific point to be removed from the data set. In general, we can identify the observations with positive and negative DFBETAs. The changes in beta estimates are relatively small, which indicates these observations don't have a meaningful impact in this case, so I don't see anything particularly influential.

4.3 Corrections for violation of the PH Assumption

```
d.model.inter1 <- coxph(Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx + obstruct*l

d.model.inter2 <- coxph(Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx + differ*log

anova(d.model.inter2)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##                    loglik     Chisq Df Pr(>|Chi|)
## NULL              -2847.5
## obstruct          -2845.2    4.7665  1  0.0290194 *
## differ            -2839.2   11.8950  1  0.0005628 ***
## extent            -2828.8   20.7393  1  5.262e-06 ***
## surg              -2826.7    4.3491  1  0.0370281 *
## node4             -2787.9   77.5495  1  < 2.2e-16 ***
## rx                -2783.8    8.0542  1  0.0045399 **
## differ:log(time) -1851.3 1865.1561  1  < 2.2e-16 ***
## extent:log(time) -1554.1  594.4044  1  < 2.2e-16 ***
## surg:log(time)   -1546.7   14.7489  1  0.0001228 ***
## rx:log(time)     -1521.2   50.9172  1  9.635e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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), except for covariate 'obstruct'. Therefore, we can include these interactions with function of time as the method of PH assumption verification and solution to its violation. Moreover, based on the results of the Analysis of Deviance procedure, the interaction of 'node4' with 'log(time)' (p-val > 0.05) is excluded from the final Cox proportional model.

4.3 Final model

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

```
d.final.model <- coxph(Surv(time, status) ~ obstruct + differ + extent + surg + node4 + rx + differ*log
summary(d.final.model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ obstruct + differ + extent +
##     surg + node4 + rx + differ * log(time) + extent * log(time) +
##     surg * log(time) + rx * log(time) - log(time), data = colon.death)
##
##   n= 906, number of events= 441
##
##                         coef  exp(coef)   se(coef)       z Pr(>|z|)
## obstruct           4.033e-01  1.497e+00  1.275e-01   3.162 0.001565 **
## differ             1.201e+01  1.650e+05  9.046e-01  13.281  < 2e-16 ***
## extent             1.745e+01  3.772e+07  8.288e-01  21.048  < 2e-16 ***
## surg               4.440e+00  8.480e+01  1.241e+00   3.578 0.000347 ***
## node4              4.114e-01  1.509e+00  1.085e-01   3.790 0.000150 ***
## rx                 4.145e+00  6.314e+01  6.293e-01   6.587 4.49e-11 ***
## differ:log(time) -1.775e+00  1.695e-01  1.333e-01 -13.319  < 2e-16 ***
## extent:log(time) -2.454e+00  8.594e-02  1.174e-01 -20.911  < 2e-16 ***
## surg:log(time)   -6.237e-01  5.359e-01  1.837e-01  -3.396 0.000684 ***
## rx:log(time)     -6.240e-01  5.358e-01  9.209e-02  -6.776 1.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## obstruct         1.497e+00  6.681e-01 1.166e+00 1.922e+00
## differ           1.650e+05  6.061e-06 2.802e+04 9.715e+05
## extent           3.772e+07  2.651e-08 7.430e+06 1.914e+08
## surg             8.480e+01  1.179e-02 7.447e+00 9.657e+02
## node4            1.509e+00  6.627e-01 1.220e+00 1.867e+00
## rx               6.314e+01  1.584e-02 1.839e+01 2.168e+02
## differ:log(time) 1.695e-01  5.901e+00 1.305e-01 2.201e-01
## extent:log(time) 8.594e-02  1.164e+01 6.828e-02 1.082e-01
## surg:log(time)   5.359e-01  1.866e+00 3.739e-01 7.682e-01
## rx:log(time)     5.358e-01  1.866e+00 4.473e-01 6.418e-01
##
## Concordance= 0.983  (se = 0.002 )
## Likelihood ratio test= 2653  on 10 df,   p=<2e-16
## Wald test            = 804  on 10 df,   p=<2e-16
## Score (logrank) test = 2513  on 10 df,   p=<2e-16
```

```
# output hazard ratio
colon.death.inter <- colon.death %>%
  mutate(rxt = rx*log(time),
         differt = differ*log(time),
         extentt = extent*log(time),
         surgt = surg*log(time))
covariates <- c("obstruct", "differ", "extent", "surg", "node4", "rx", "differt", "extentt", "surgt", "
```

```
univ_formulas <- sapply(covariates,
                        function(x) as.formula(paste('Surv(time, status)~', x)))

univ_models <- lapply(univ_formulas, function(x){coxph(x, data = colon.death.inter)})
# Extract data
univ_results <- lapply(univ_models,
                       function(x){
                          x <- summary(x)
                          p.value<-signif(x$wald["pvalue"], digits=2)
                          wald.test<-signif(x$wald["test"], digits=2)
                          beta<-signif(x$coef[1], digits=2);#coeficient beta
                          HR <-signif(x$coef[2], digits=2);#exp(beta)
                          HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                          HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                          HR <- paste0(HR, " (",
                                       HR.confint.lower, "-", HR.confint.upper, ")")
                          res<-c(beta, HR, wald.test, p.value)
                          names(res)<-c("beta", "HR (95% CI for HR)", "wald.test",
                                        "p.value")
                          return(res)
                          #return(exp(cbind(coef(x),confint(x))))
                         })
res <- t(as.data.frame(univ_results, check.names = FALSE))
as.data.frame(res)
```

```
##            beta HR (95% CI for HR) wald.test p.value
## obstruct   0.26        1.3 (1-1.6)         5   0.025
## differ     0.33      1.4 (1.1-1.7)        12 0.00065
## extent     0.53      1.7 (1.4-2.1)        23   2e-06
## surg       0.23        1.3 (1-1.5)       4.9   0.027
## node4      0.95      2.6 (2.1-3.1)        94 2.9e-22
## rx        -0.17   0.84 (0.75-0.94)         9  0.0027
## differt   -0.11    0.9 (0.88-0.92)        78 1.3e-18
## extentt    -0.1    0.9 (0.89-0.92)       120 6.4e-28
## surgt    0.0037         1 (0.98-1)      0.07     0.8
## rxt      -0.074   0.93 (0.91-0.94)        83 8.4e-20
```

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 rxLev+5Fu and the observation group is 0.678. Furthermore, the confidence interval for the hazard ratio is (0.55, 0.89). The coefficients in a Cox regression relate to hazard; a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated.

When controlling the age at transplant in this time-varying covariate model, the hazard rate of death for those who did not receive transplant is 5.965 times greater than that for those who received transplant, with p-value <.0001. We can conclude that there is a significant survival difference between transplant or not.