PSTAT175-ProjectBrief

Kartik Adimulam

2025-04-29

```r
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2

## Loading required package: ggpubr

##
## Attaching package: 'survminer'

## The following object is masked from 'package:survival':
##
##     myeloma
```

```r
library(ggplot2)
library(tidyr)
library(xfun)
```

```
##
## Attaching package: 'xfun'

## The following objects are masked from 'package:base':
##
##     attr, isFALSE
```

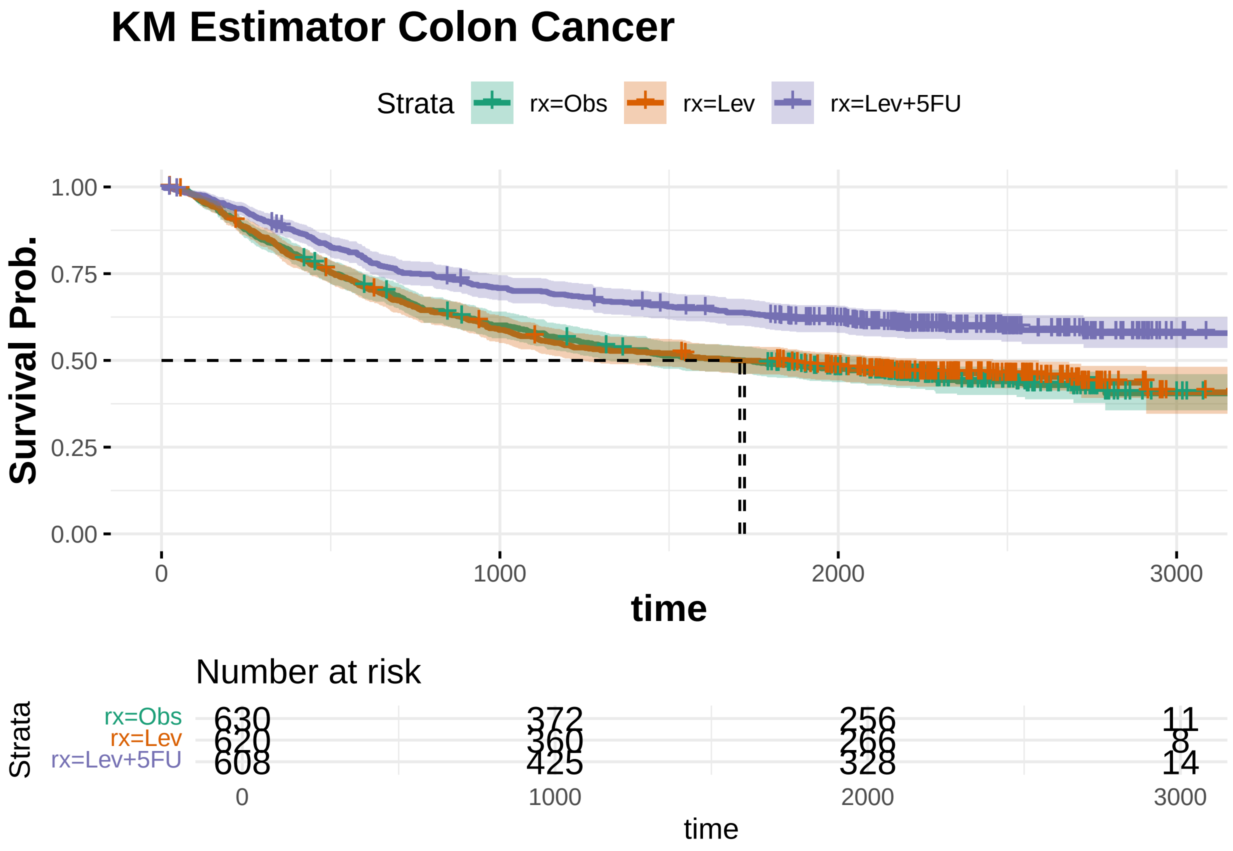
```r
head(colon)
```

```
##   id study      rx sex age obstruct perfor adhere nodes status differ extent
## 1  1     1 Lev+5FU   1  43        0      0      0     5      1      2      3
## 2  1     1 Lev+5FU   1  43        0      0      0     5      1      2      3
## 3  2     1 Lev+5FU   1  63        0      0      0     1      0      2      3
## 4  2     1 Lev+5FU   1  63        0      0      0     1      0      2      3
## 5  3     1     Obs   0  71        0      0      1     7      1      2      2
## 6  3     1     Obs   0  71        0      0      1     7      1      2      2
##   surg node4 time etype
## 1    0     1 1521     2
## 2    0     1  968     1
## 3    0     0 3087     2
## 4    0     0 3087     1
## 5    0     1  963     2
## 6    0     1  542     1
```

```
colon=colon
colon_fit = survfit(Surv(time, status) ~ rx, colon)

ggsurvplot(colon_fit, colon, title='KM Estimator Colon Cancer',xlab='time',
           ylab='Survival Prob.',surv.median.line = 'hv',palette = "Dark2",
           conf.int = TRUE,
           risk.table = TRUE,
           ggtheme = theme_minimal(),
           font.main = c(16, "bold", "black"),
           font.x = c(14, "bold"),
           font.y = c(14, "bold"))
```

```
## Warning in geom_segment(aes(x = 0, y = max(y2), xend = max(x1), yend = max(y2)), : All aesthetics ha
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
## All aesthetics have length 1, but the data has 2 rows.
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
## All aesthetics have length 1, but the data has 2 rows.
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
## All aesthetics have length 1, but the data has 2 rows.
## i Please consider using ‘annotate()‘ or provide this layer with data containing
##   a single row.
```



Overview of Dataset

1. Our dataset contains survival time data of days until death of colon cancer patients that received different levels of chemotherapy. The primary covariate in our analysis is the ‘rx’ treatment variable which has 3 levels (No treatment, low-toxicity treatment, and moderate toxicity treatment) and the censoring status indicator. There are other covariates such as sex, age, obstruct(was the colon obstructed by the tumor?), perfor (was the colon perforated?), adhere (is the tumor on nearby organs?), nodes (how many lymph nodes have cancer?), differ (is the tumor differentiated?), extent (how much has the tumor spread locally?), surg (how long from surgery to registration?), node4 (indicator variable of more than 4 positive lymph nodes), and etype (did subject’s cancer recur or death). The scientific questions we are trying to answer are: how do these covariates affect the probability of surival? are there some covariates that are not strong predictors and if so which ones? How much is survival probability affected by different covariate values? Do coviarates affect survival probability differently over time? Are there important interactions between these covariates?

Bibliography

Laurie, J A et al. “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.” Journal of clinical oncology : official journal of the American Society of Clinical Oncology vol. 7,10 (1989): 1447-56. doi:10.1200/JCO.1989.7.10.1447

Splitting Data

```
colon.recur = colon[colon$etype==1,]
colon.death = colon[colon$etype==2,]

colon.recur.cox = coxph(Surv(time,status)~rx, colon.recur)
summary(colon.recur.cox)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx, data = colon.recur)
##
##   n= 929, number of events= 468
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev      -0.01512   0.98499  0.10708 -0.141    0.888
## rxLev+5FU -0.51209   0.59924  0.11863 -4.317 1.58e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9850      1.015    0.7985    1.2150
## rxLev+5FU    0.5992      1.669    0.4749    0.7561
##
## Concordance= 0.554  (se = 0.013 )
## Likelihood ratio test= 24.34  on 2 df,   p=5e-06
## Wald test            = 22.58  on 2 df,   p=1e-05
## Score (logrank) test = 23.07  on 2 df,   p=1e-05
```

```r
anova(colon.recur.cox)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -3040.3
## rx   -3028.1 24.343  2  5.175e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
colon.death.cox = coxph(Surv(time, status) ~ rx, colon.death)
summary(colon.death.cox)
```

```
## 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
```

```r
anova(colon.death.cox)
```

```
## 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
## rx   -2924.1 12.148  2   0.002302 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Low toxicity treatment doesn’t have a significant effect on recurrence or death, but moderately toxic treatment does. Overall the rx covariate is significant on time until recurrence and death.

```
colon.recur.CI = exp(confint(colon.recur.cox))
colon.death.CI = exp(confint(colon.death.cox))
```

Again the CIs show that Lev is not significant but Lev+5FU is significant in reducing time until recurrence and death. Our hazard ratio CI for Lev treatment contains 1 but our hazard ratio CI for Lev+5FU treatment does not. We are 95% confident that Lev+5FU treatment significantly affects time until recurrence.

```
exp(coef(colon.recur.cox))
```

```
##     rxLev rxLev+5FU
## 0.9849905 0.5992400
```

```
exp(coef(colon.death.cox))
```

```
##     rxLev rxLev+5FU
## 0.9737142 0.6895540
```

At any point in time Lev+5FU treated patients are 0.6 times as likely to have a colon cancer recurrence compared to patients without treatment. The patients with Lev+5FU are 0.69 times as likely to have a colon cancer death compared to patients without treatment.

Model Fitting

Recurrence of Cancer Subset

Using forward stepwise selection and AIC, we can find the most relevant covariates to build and finds the best model at predicting survival times. We first perform this process on the subset of data where there was a recurrence of cancer in patients.

```
colon.rx.sex = coxph(Surv(time,status)~rx + sex, colon.recur)
colon.rx.age = coxph(Surv(time,status)~rx + age, colon.recur)
colon.rx.obstruct = coxph(Surv(time,status)~rx + obstruct, colon.recur)
colon.rx.perfor = coxph(Surv(time,status)~rx + perfor, colon.recur)
colon.rx.adhere = coxph(Surv(time,status)~rx + adhere, colon.recur)
colon.rx.nodes = coxph(Surv(time,status)~rx + nodes, colon.recur)
colon.rx.differ = coxph(Surv(time,status)~rx + differ, colon.recur)
colon.rx.extent = coxph(Surv(time,status)~rx + extent, colon.recur)
colon.rx.surg = coxph(Surv(time,status)~rx + surg, colon.recur)
colon.rx.node4 = coxph(Surv(time,status)~rx + node4, colon.recur)

AIC(colon.rx.sex, colon.rx.age, colon.rx.obstruct, colon.rx.perfor,
    colon.rx.adhere, colon.rx.nodes, colon.rx.differ, colon.rx.extent,
    colon.rx.surg, colon.rx.node4)
```

```
## Warning in AIC.default(colon.rx.sex, colon.rx.age, colon.rx.obstruct,
## colon.rx.perfor, : models are not all fitted to the same number of observations
```

```
##                   df      AIC
## colon.rx.sex       3 6060.887
## colon.rx.age       3 6059.623
## colon.rx.obstruct  3 6058.401
## colon.rx.perfor    3 6060.226
## colon.rx.adhere    3 6056.852
## colon.rx.nodes     3 5827.396
## colon.rx.differ    3 5898.618
## colon.rx.extent    3 6030.710
## colon.rx.surg      3 6056.936
## colon.rx.node4     3 5983.621
```

The best model is adding nodes as a covariate, attaining an AIC of 5827.4. The second best model is differentiation of tumor and the third best is node4 (more than 4 positive lymph nodes). However, we should not include node4 since it is highly correlated with nodes. They are essentially measuring the same thing.

```
colon.rx.nodes.differ = coxph(Surv(time,status)~rx + nodes + differ, colon.recur)
colon.rx.nodes.sex = coxph(Surv(time,status)~rx + nodes + sex, colon.recur)
colon.rx.nodes.age = coxph(Surv(time,status)~rx + nodes + age, colon.recur)
colon.rx.nodes.obstruct = coxph(Surv(time,status)~rx + nodes + obstruct, colon.recur)
colon.rx.nodes.perfor = coxph(Surv(time,status)~rx + nodes + perfor, colon.recur)
colon.rx.nodes.adhere = coxph(Surv(time,status)~rx + nodes + adhere, colon.recur)
colon.rx.nodes.extent = coxph(Surv(time,status)~rx + nodes + extent, colon.recur)
colon.rx.nodes.surg = coxph(Surv(time,status)~rx + nodes + surg, colon.recur)

AIC(colon.rx.nodes.differ, colon.rx.nodes.sex, colon.rx.nodes.age, colon.rx.nodes.obstruct, colon.rx.noc
```

```
## Warning in AIC.default(colon.rx.nodes.differ, colon.rx.nodes.sex,
## colon.rx.nodes.age, : models are not all fitted to the same number of
## observations
```

```
##                         df      AIC
## colon.rx.nodes.differ    4 5676.214
## colon.rx.nodes.sex       4 5826.886
## colon.rx.nodes.age       4 5828.648
## colon.rx.nodes.obstruct  4 5824.943
## colon.rx.nodes.perfor    4 5827.301
## colon.rx.nodes.adhere    4 5824.139
## colon.rx.nodes.extent    4 5802.637
## colon.rx.nodes.surg      4 5824.080
```

Adding differ as the next covariate is the best model, as our model now has an AIC of 5676.21

```
colon.rnd.sex = coxph(Surv(time,status)~rx + nodes + differ + sex, colon.recur)
colon.rnd.age  = coxph(Surv(time,status)~rx + nodes + differ + age, colon.recur)
colon.rnd.obstruct  = coxph(Surv(time,status)~rx + nodes + differ + obstruct, colon.recur)
colon.rnd.perfor  = coxph(Surv(time,status)~rx + nodes + differ +perfor , colon.recur)
colon.rnd.adhere = coxph(Surv(time,status)~rx + nodes + differ +adhere , colon.recur)
colon.rnd.extent = coxph(Surv(time,status)~rx + nodes + differ + extent, colon.recur)
colon.rnd.surg  = coxph(Surv(time,status)~rx + nodes + differ +surg , colon.recur)
AIC(colon.rnd.sex, colon.rnd.age, colon.rnd.obstruct, colon.rnd.perfor, colon.rnd.adhere, colon.rnd.exte
```

```
##                    df      AIC
## colon.rnd.sex       5 5675.069
## colon.rnd.age       5 5677.614
## colon.rnd.obstruct  5 5673.698
## colon.rnd.perfor    5 5676.290
## colon.rnd.adhere    5 5674.507
## colon.rnd.extent    5 5658.703
## colon.rnd.surg      5 5673.328
```

Adding extent as the next covariate is the best model, attaining an AIC of 5658.7

```
BIC(colon.rnd.extent,colon.rx.nodes.differ, colon.rx.nodes, colon.recur.cox)
```

```
## Warning in BIC.default(colon.rnd.extent, colon.rx.nodes.differ, colon.rx.nodes,
## : models are not all fitted to the same number of observations
```

```
##                       df      BIC
## colon.rnd.extent       5 5679.205
## colon.rx.nodes.differ  4 5692.615
## colon.rx.nodes         3 5839.763
## colon.recur.cox        2 6068.505
```

Using BIC as a sanity check helps us to verify that adding extent as another covariate is still beneficial to our model's prediction accuracy while avoiding overfitting (too many covariates), attaining a BIC of 5679.21.

```
colon.rnde.sex = coxph(Surv(time,status)~rx + nodes + differ + extent + sex, colon.recur)
colon.rnde.age = coxph(Surv(time,status)~rx + nodes + differ + extent + age, colon.recur)
colon.rnde.obstruct = coxph(Surv(time,status)~rx + nodes + differ + extent + obstruct, colon.recur)
colon.rnde.perfor = coxph(Surv(time,status)~rx + nodes + differ + extent + perfor, colon.recur)
colon.rnde.adhere = coxph(Surv(time,status)~rx + nodes + differ + extent + adhere, colon.recur)
colon.rnde.surg = coxph(Surv(time,status)~rx + nodes + differ + extent + surg, colon.recur)

AIC(colon.rnde.sex, colon.rnde.age, colon.rnde.obstruct, colon.rnde.perfor,
    colon.rnde.adhere, colon.rnde.surg)
```

```
##                     df      AIC
## colon.rnde.sex       6 5657.874
## colon.rnde.age       6 5660.177
## colon.rnde.obstruct  6 5657.612
## colon.rnde.perfor    6 5659.534
## colon.rnde.adhere    6 5658.749
## colon.rnde.surg      6 5655.546
```

```
BIC(colon.rnde.sex, colon.rnde.age, colon.rnde.obstruct, colon.rnde.perfor,
    colon.rnde.adhere, colon.rnde.surg)
```

```
##                     df      BIC
## colon.rnde.sex       6 5682.476
## colon.rnde.age       6 5684.779
## colon.rnde.obstruct  6 5682.214
## colon.rnde.perfor    6 5684.136
## colon.rnde.adhere    6 5683.351
## colon.rnde.surg      6 5680.148
```

Adding Surg (time from surgery until registration into the study) as our next covariate again lowers our AIC to 5655.55

```
colon.rndes.sex = coxph(Surv(time,status)~rx + nodes + differ + extent + surg + sex, colon.recur)
colon.rndes.age = coxph(Surv(time,status)~rx + nodes + differ + extent + surg + age, colon.recur)
colon.rndes.obstruct = coxph(Surv(time,status)~rx + nodes + differ + extent + surg + obstruct, colon.re
colon.rndes.perfor = coxph(Surv(time,status)~rx + nodes + differ + extent + surg + perfor, colon.recur)
colon.rndes.adhere = coxph(Surv(time,status)~rx + nodes + differ + extent + surg + adhere, colon.recur)

AIC(colon.rndes.sex, colon.rndes.age, colon.rndes.obstruct, colon.rndes.perfor, colon.rndes.adhere)
```

```
##                      df      AIC
## colon.rndes.sex       7 5654.365
## colon.rndes.age       7 5656.926
## colon.rndes.obstruct  7 5654.744
## colon.rndes.perfor    7 5656.414
## colon.rndes.adhere    7 5655.639
```

We attain a slightly lower AIC of 5654.37 by adding sex as our next covariate. However, since the change is small we will verify this using BIC score too.

```
BIC(colon.rndes.sex, colon.rnde.surg)
```

```
##                 df      BIC
## colon.rndes.sex  7 5683.067
## colon.rnde.surg  6 5680.148
```

Using BIC score instead, which penalizes more covariates more intensely than AIC, we see that our previous model is actually preferable, so we will stop adding covariates at Surg.

```
anova(colon.rndes.sex)
```

```
## 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
## rx     -2865.6 23.1505  2  9.396e-06 ***
## nodes  -2835.8 59.6074  1  1.158e-14 ***
## differ -2834.1  3.2934  1    0.06956 .
## extent -2824.3 19.5108  1  1.000e-05 ***
## surg   -2821.8  5.1569  1    0.02315 *
## sex    -2820.2  3.1816  1    0.07447 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

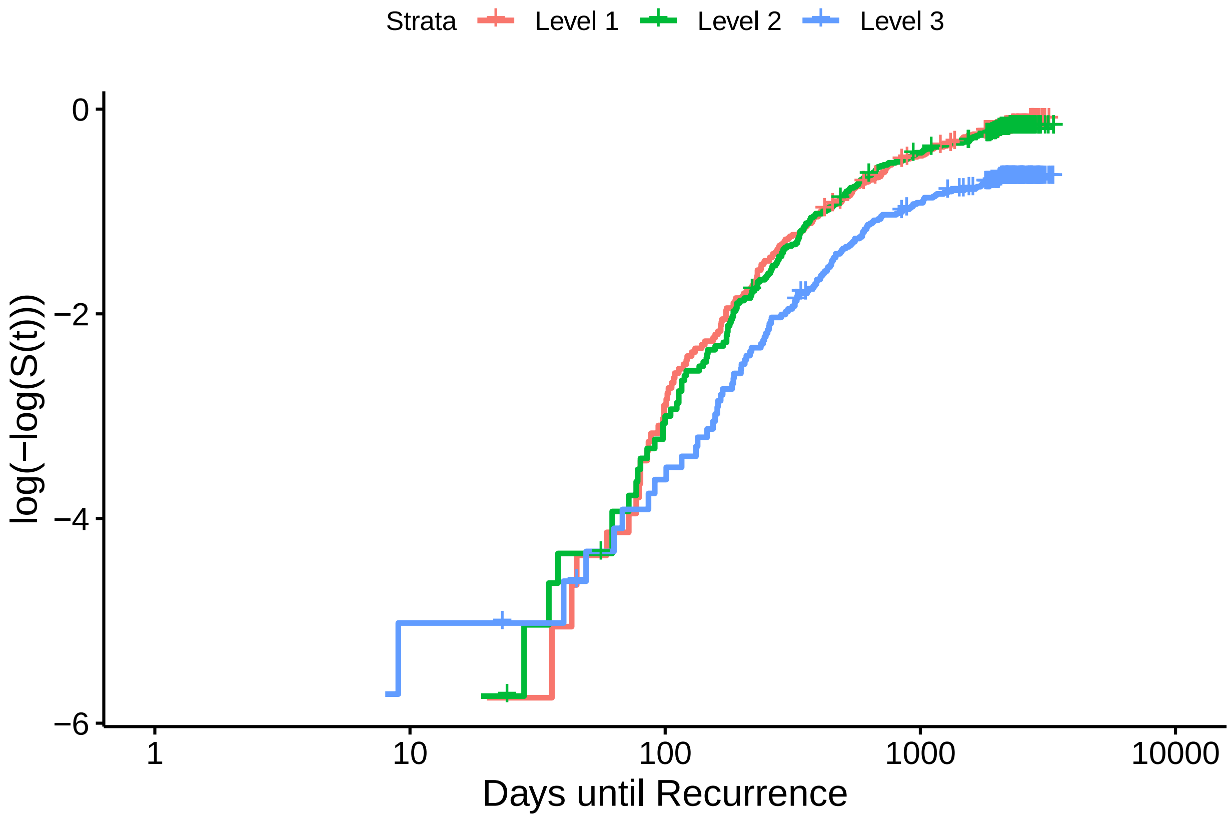
We also verify this by checking that if we add sex as a covariate, it is not significantly affecting time until recurrence from the likelihood ratio test.

Death Subset

Using the same exact process as above for the data with etype=1 (recurrence of cancer), we find that for etype = 2 (death), the model had the same covariates until adding the 6th covariate. Adding obstruct to the death cox model gave us the next best model which is different from the recur data, where at this point, adding sex as our next covariate gave us the lower AIC score. However, in both subsets of data, adding the 6th covariate gave us a higher BIC score, allowing us to leave it out and have the same covariates (rx, node, differ, extent, surg).

Checking Proportional Hazards Assumptions

```
colon.fit = survfit(Surv(time, status) ~ rx, colon.recur)
ggsurvplot(colon.fit, colon.recur, legend.labs=c('Level 1','Level 2', 'Level 3'), fun='cloglog') + labs
```



```
cox.zph(colon.recur.cox)
```

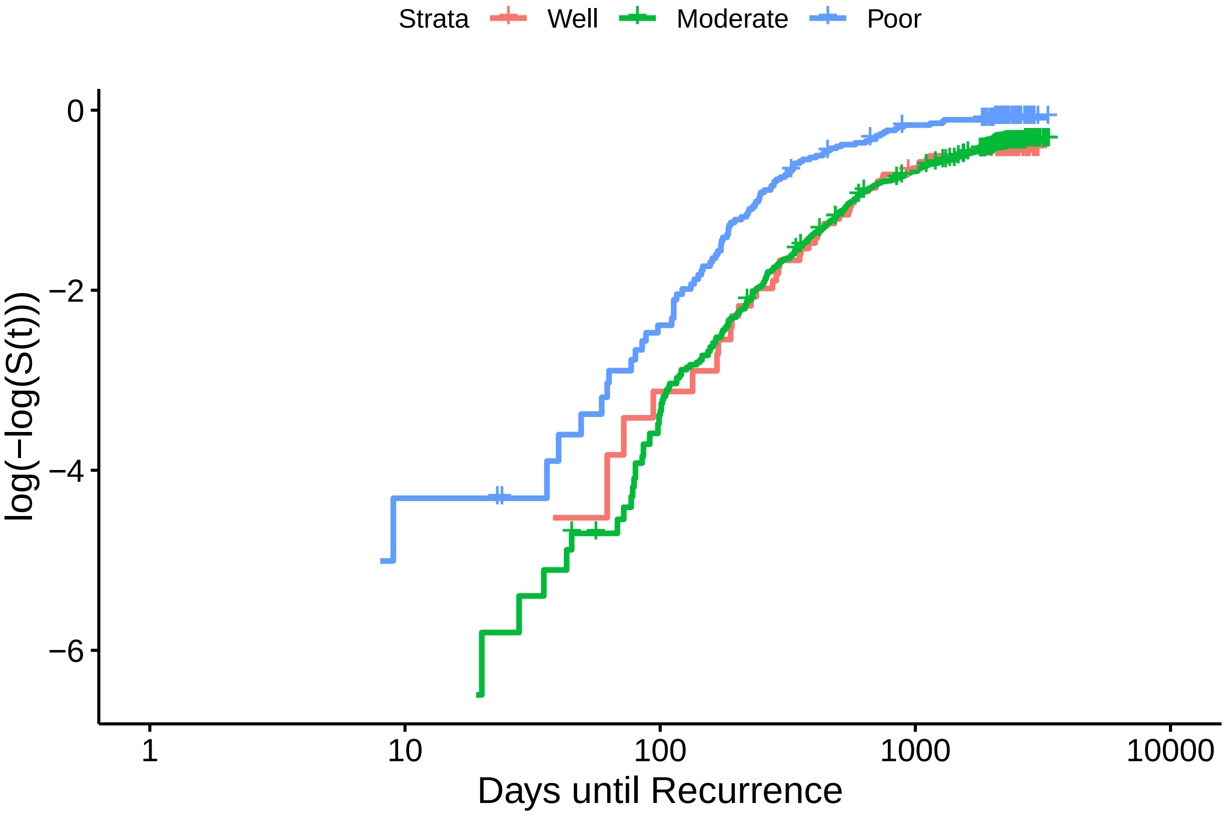
```
##        chisq df    p
## rx     0.301  2 0.86
## GLOBAL 0.301  2 0.86
```

The proportional hazards assumption seems violated from the log-log plot for Baseline and low toxicity treatment, but we have already observed that those two treatments do not significantly differ in hazard rates. It appears that the difference between low toxicity and medium toxicity does not violate the proportional hazards assumption, and we verify this using the Schoenfield residuals test attaining a p-value of 0.86 for the rx covariate.

```
cox.zph(colon.rnde.surg)
```

```
##          chisq df       p
## rx       0.454  2 0.79696
## nodes    1.295  1 0.25504
## differ 13.525  1 0.00024
## extent   0.140  1 0.70802
## surg     1.827  1 0.17644
## GLOBAL 16.112  6 0.01316
```

```
colon.fit.differ = survfit(Surv(time, status) ~ differ, colon.recur)
ggsurvplot(colon.fit.differ, colon.recur, legend.labs=c('Well','Moderate', 'Poor'), fun='cloglog') + lab
```



From our cox ZPH test we see that differ could potentially be violating the proportional hazards assumption with a p-value of 0.00024 and we verify this using a log-log plot. The assumption seems to hold for the difference between ‘poor’ and ‘well’ and ‘moderate’ but not for the difference between ‘well’ and ‘moderate’.