PSTAT175-ProjectBrief

Kartik Adimulam

2025-04-29

1. Introduction

Overview of the Dataset

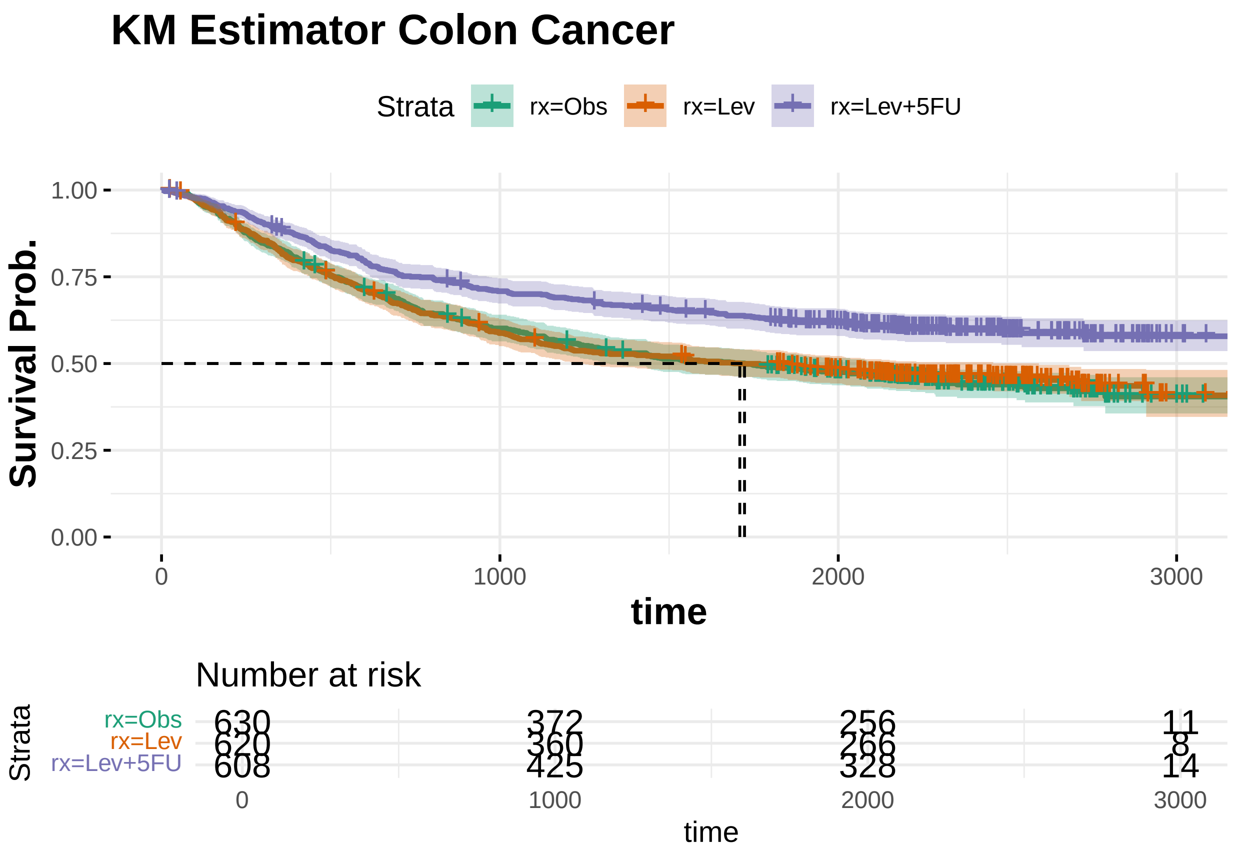
Our dataset contains survival time data of days until death of colon cancer patients that received different levels of chemotherapy. The primary covariate in our analysis is the 'rx' treatment variable which has 3 levels (No treatment, low-toxicity treatment, and moderate toxicity treatment) and the censoring status indicator. There are other covariates such as:
- sex: male (1) or female (0)
- age: in years
- 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
- etype: did subject's cancer recur or death

Motivating Questions

The scientific questions we are trying to answer are:
1. Does the type of treatment affect survival times?
2. Are there some covariates that are not strong predictors and if so which ones?
3. How much is survival probability affected by different covariate values?
4. Do covariates affect survival probability differently over time?
5. Are there important interactions between these covariates?

Survival Probability of Treatments

```
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"))
```



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
```

```
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
```

```
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
```

```
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
```

In our time until recurrence model, treatment using Levamisole compared with the baseline effect has a p-value of 0.888. This teaches us that this low-toxicity compound is not effective in reducing the time until recurrence for colon cancer patients.

However, the p-value of the treatment which combines 5FU (moderately toxic) treatment and Levamisole (low toxicitity) is 0.00002. We can conclude that the combination of 5FU and Levamisole is effective in reducing time until recurrence but the treatment of just using Levamisole is not.

Overall, we observe using a likelihood ratio test p-value of ~0 that the treatment itself is effective in reducing time until recurrence, but most of this effect is likely due to the high effectiveness of Levamisole+5FU treatment.

In our time until death model, treatment using Levamisole compared with the baseline effect has a p-value of 0.809. This teaches us that this low-toxicity compound is not effective in reducing the time until death for colon cancer patients.

However, the p-value of the treatment which combines 5FU (moderately toxic) treatment and Levamisole (low toxicitity) is 0.0018. We can conclude that the combination of 5FU and Levamisole is effective in reducing time until death but the treatment of just using Levamisole is not.

Overall, we observe using a likelihood ratio test p-value of 0.0023 that the treatment itself is effective in reducing time until death, but most of this effect is likely due to the high effectiveness of Levamisole+5FU treatment.

Confidence Intervals

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

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

Hazard Ratio

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

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

```r
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.

2. 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)
```

```
##                   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
```

```
##                         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)
```

```
##                       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.

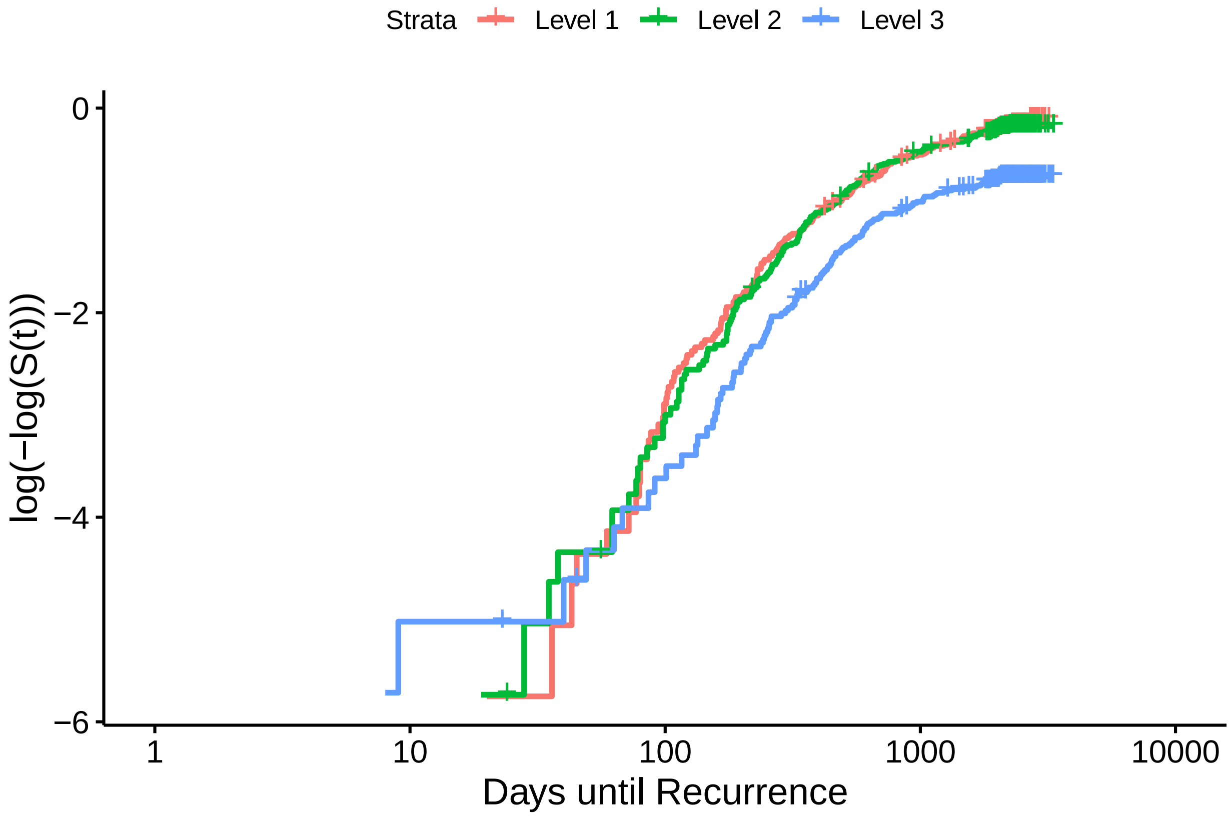
```
anova(colon.rndes.sex)
```

```
## 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.

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(x='Days until Recurrence')
```



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

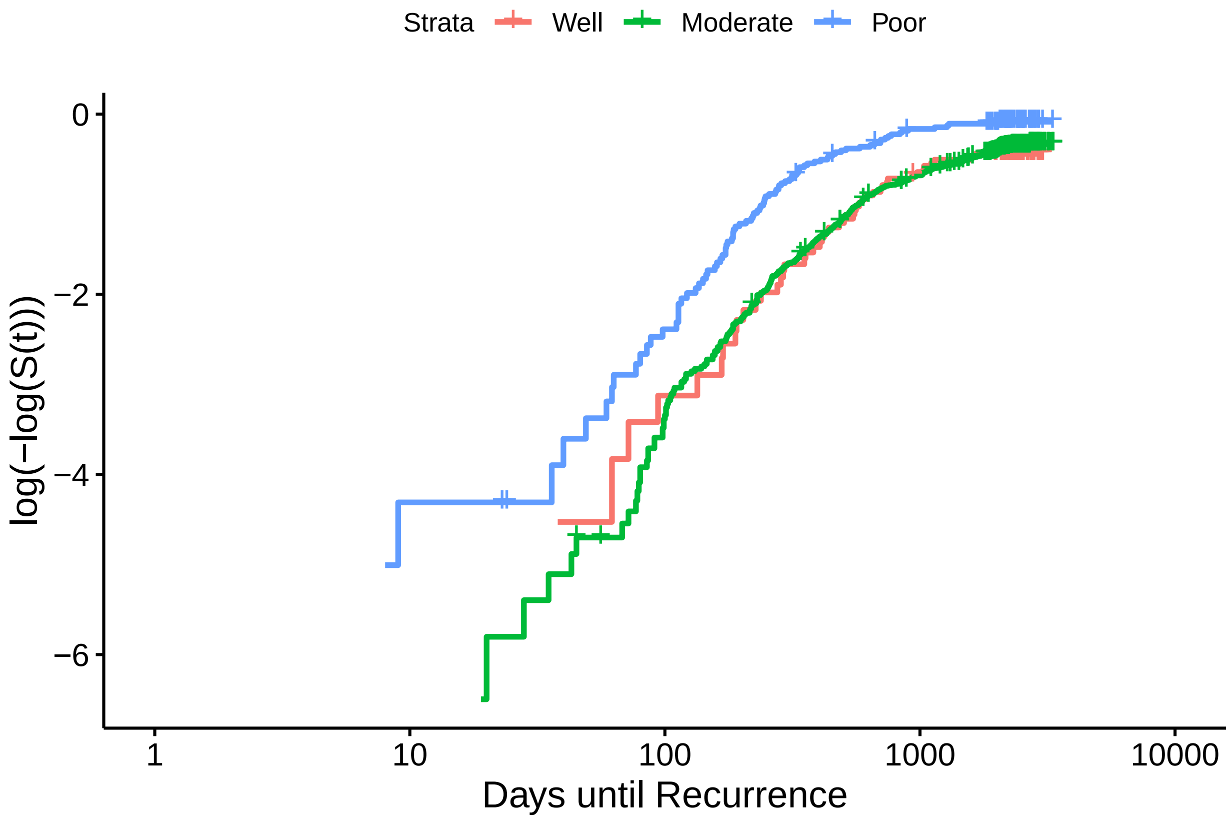
```
##        chisq df    p
## rx     0.301  2 0.86
## GLOBAL 0.301  2 0.86
```

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
```

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



From our cox ZPH test we see that the differentiation of tumor covariate violates the proportional hazards assumption with a p-value of 0.00024 and we verify this using a log-log plot. We can observe many cross-over points between the plot for ‘moderate’ differentiation and ‘well’ differentiation. To account for this violation, we will attempt to stratify on the differentiation covariate and produce a stratified Cox PH model.

```r
colon.rnde.surg = coxph(Surv(time,status)~rx + nodes + strata(differ) + extent + surg, colon.recur)
anova(colon.rnde.surg)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##         loglik   Chisq Df Pr(>|Chi|)
## NULL   -2529.3
## rx     -2517.6 23.3621  2  8.453e-06 ***
## nodes  -2491.5 52.1831  1  5.056e-13 ***
## extent -2481.4 20.1808  1  7.046e-06 ***
## surg   -2478.7  5.4127  1    0.01999 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##           chisq df    p
## rx     0.76738  2 0.68
## nodes  0.25978  1 0.61
## extent 0.00222  1 0.96
## surg   1.70609  1 0.19
## GLOBAL 2.61561  5 0.76
```

Our AIC drops to 4967.46 after stratifying on the differentiation variable instead of simply including it as a covariate in our model. This is a significant decrease in AIC most likely due to the fact that differentation should have never been included as a covariate in the model, as it violated the proportional hazards assumption.

To verify again that none of our covariates violate this assumption, we will run the Cox ZPH test on our new stratified model. We observe that all of our p-values are greater than 0.05 and we can keep them in our model as covariates.

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).

The best model in the data for patients with a recorded death adds nodes with an AIC of 5620.639. The second best model adds differ with an AIC of 5678.121 and the third best model adds node4 with an AIC of 5764.921.

After adding differ we got an AIC of 5458.958.

Adding extent give us an AIC of 5441.828.

Adding surg gives an AIC of 5438.274.

Adding obstruct gives us an AIC of 5436.907, which is different from the recur data, where at this point, adding sex as out next covariate gave us the lower AIC score.

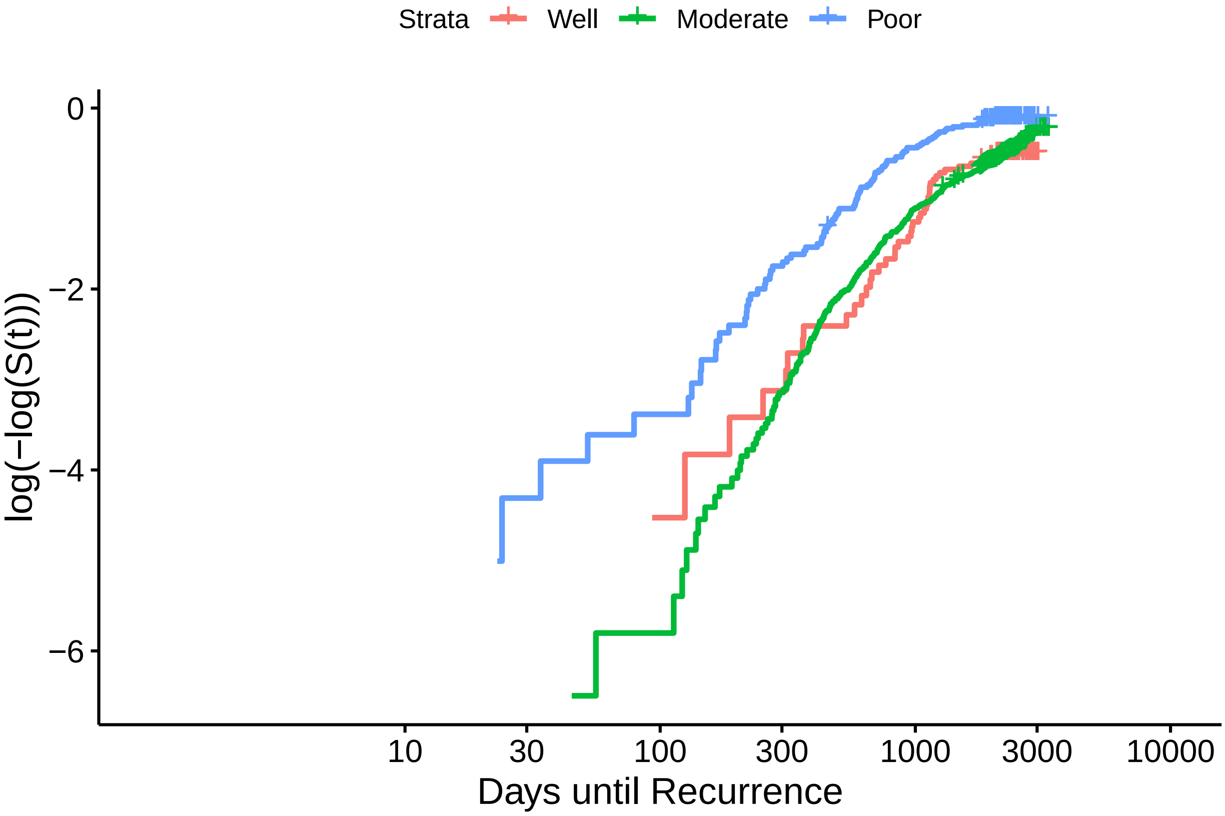
By using BIC, which penalizes the number of covariates more than AIC, we can see that the previous model, without obstruct, is a better model as it gives a BIC of 5462.657. Therefore, we will use the exact same covariates as the recurrence model.

Checking Proportional Hazards Assumptions

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

```
##          chisq df       p
## rx      2.6512  2 0.26565
## nodes   0.0206  1 0.88595
## differ 12.3575  1 0.00044
## extent  3.5707  1 0.05881
## surg    0.0213  1 0.88386
## GLOBAL 18.4058  6 0.00529
```

```
colon.fit.differ = survfit(Surv(time, status) ~ differ, colon.death)
ggsurvplot(colon.fit.differ, colon.death, legend.labs=c('Well','Moderate', 'Poor'), fun='cloglog') +
  labs(x='Days until Recurrence')
```



We can also observe, that similarly to our recurrence model the differentiation covariate again violates the PH assumption. We can verify this using a log-log plot, which shows many cross-over points bewteen 'moderate' differentiation and 'well' differentiation. We will remedy this using a stratification on the differentiation variable.

```
death.rnde.surg =  coxph(Surv(time,status)~rx + nodes + strata(differ) + extent + surg, colon.death)
AIC(death.rnde.surg)
```

```
## [1] 4769.913
```

Our AIC drops to 4769.91 after stratifying on the differentiation covariate instead of simply using it as a covariate in our model. This is a significant decrease and proves that our previous model had an alarming violation of the PH assumption.

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

```
##         chisq df    p
## rx     2.2824  2 0.32
## nodes  0.7298  1 0.39
## extent 2.6415  1 0.10
## surg   0.0332  1 0.86
## GLOBAL 6.2753  5 0.28
```

We run cox ZPH again on this new stratified cox PH model to ensure no violations and observe that we have successfully remedied our issue. All of our p-values from the cox ZPH test are above 0.05.

3. Advanced Models

Recurrent Events Model

```r
for(i in 1:1858){
  ifelse(colon$etype[i]==2, colon$start[i] <- colon$time[i+1], colon$start[i] <- 0 ) }
colon$stop = colon$time
colon$stop = ifelse(colon$stop<=colon$start, colon$stop<-colon$stop+0.0001, colon$stop<-colon$stop)

colon.cox = coxph(Surv(start,stop,status)~rx, colon)
summary(colon.cox)
```

```
## Call:
## coxph(formula = Surv(start, stop, status) ~ rx, data = colon)
##
##   n= 1858, number of events= 920
##
##                coef exp(coef) se(coef)      z Pr(>|z|)
## rxLev     -0.01672   0.98342  0.07683 -0.218    0.828
## rxLev+5FU -0.38646   0.67946  0.08390 -4.606  4.1e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev        0.9834      1.017    0.8459    1.1432
## rxLev+5FU    0.6795      1.472    0.5764    0.8009
##
## Concordance= 0.539  (se = 0.01 )
## Likelihood ratio test= 26.98  on 2 df,   p=1e-06
## Wald test            = 25.53  on 2 df,   p=3e-06
## Score (logrank) test = 25.83  on 2 df,   p=2e-06
```

In our recurrent events model, we again observe that the Lemavisole treatment in itself is not effective in reducing time until recurrence or death. However, Levamisole treatment combined with the moderately toxic 5FU treatment is effective in reducing time until recurrence or death.

4. Citations

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