# A Survival Analysis on Standard and Chemotherapy Treatments for Lung Cancer

Luke Todd, Riley Zamora, Mac Beggs

2022-11-25

## 1 Introduction

Lung cancer is a common disease that affects millions of people per year. As time goes on, more and more efforts and research are focused on finding the cure for cancer, or at least a more effective treatment. One study was carried out by the US Veterans Administration that tested two treatments on 137 patients that had advanced inoperable lung cancer. 69 patients received the standard treatment and 68 patients received the chemotherapy treatment. The original intention of the study was to look at the difference of the standard treatment and the newer chemotherapy treatment. This dataset is included in the "survival" package for R. In our study, we will use this data to investigate whether or not the chemotherapy treatment is extending the lives of the lung cancer patients. To do so, we will utilize covariates like "time", the survival time or days until an event, "status", the censoring status, and others like Karnofsky performance score and cell type. We will utilize survival analysis techniques like Kaplan-Meier curves, Cox PH models, log-log plots, Schoenfeld residuals, time-varying covariates, and more.

## 2 Analysis

We will start the analysis of this dataset by first loading the data and looking for any obscure data points. Once the data is ready, we will plot a basic Kaplan Meier curve, as well as another Kaplan Meier curve that shows the survival function for the standard treatment and the chemotherapy treatment.

### 2.1 Loading Data

```
veteran <- survival::veteran

vet.time <- veteran$time
vet.status <- veteran$status

vet.surv <- Surv(vet.time, vet.status)

vet.survfit <- survfit(vet.surv ~ 1)
vet.survfit
```

```
## Call: survfit(formula = vet.surv ~ 1)
##
##        n events median 0.95LCL 0.95UCL
## [1,] 137    128     80      52     105
```

From this output, we can see that out of the 137 patients, 128 of them had events (died), and 9 of them were censored by the end of the experiment (alive).

Looking at the raw data, it is obvious that there are two values that is significantly larger than every other value. This value corresponds to row 70 and row 75 in "veteran", where the survival time is 999 and 991, respectively; the next closest survival time is 587. We will remove these values to more accurately model our data.

```r
veteran <- veteran[-c(70, 75), ]
```

## 2.2 Kaplan-Meier Curves

Using the cleaned up dataset, we will now plot two Kaplan Meier plots. The first figure will be a basic survival plot that includes all covariates included in the dataset. The second figure will split the dataset into two different Kaplan Meier models: one for the standard treatment and another for the chemotherapy treatement. This will just a quick visualization of how the survival rate changes over time. It will also give a preliminary insight to how different or similar of an effect that the two different treatments have on survival rate.

```r
ggsurvplot(survfit(Surv(time, status) ~ 1, data = veteran),
     xlab = "Time (in months)",
     ylab = "Survival Proportion",
     title = "Survival Proportion vs. Time",
     legend.labs = c("All TRT"),
     risk.table = TRUE,
     risk.table.col = "strata")
```

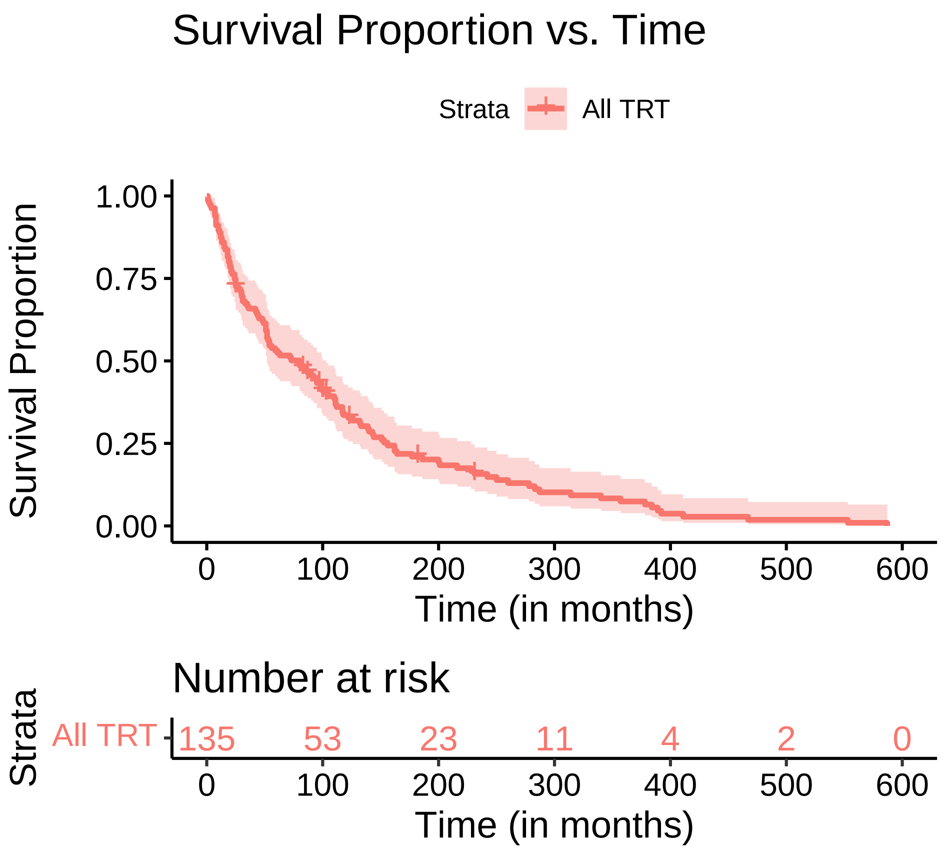
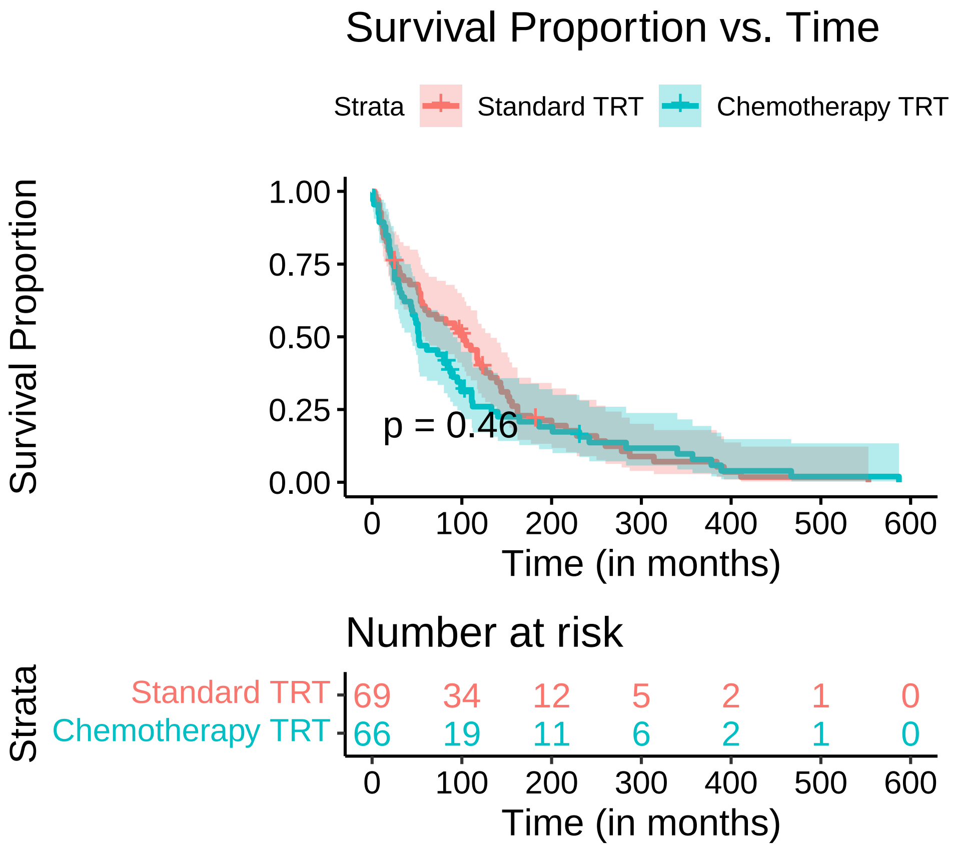


Figure 1: Kaplan-Meier curve for all covariates.

```r
ggsurvplot(survfit(Surv(time, status) ~ as.factor(trt), data = veteran),
           conf.int = TRUE,
           pval = TRUE,
           risk.table = TRUE,
           risk.table.height = 0.3,
           risk.table.col = "strata",
           legend.labs = c("Standard TRT", "Chemotherapy TRT"),
           xlab = "Time (in months)",
           ylab = "Survival Proportion",
           title = "Survival Proportion vs. Time")
```



| Strata | 0 | 100 | 200 | 300 | 400 | 500 | 600 |
|---|---|---|---|---|---|---|---|
| Standard TRT | 69 | 34 | 12 | 5 | 2 | 1 | 0 |
| Chemotherapy TRT | 66 | 19 | 11 | 6 | 2 | 1 | 0 |

Figure 2: Kaplan-Meier curve split by each treatment.

**Figure 2** gives a p-value of 0.46. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference between the standard treatment and the chemotherapy treatment. **Figure 2** gives a good look of how significant this covariate will be; we will see how this value changes over time when we build a model with only the significant covariates.

We will now move into Cox Proportional Hazard models in order to look at hazard ratios.

## 2.3 Cox PH Models

First, we will see how our covariate of interest, treatment, performs in a Cox PH model.

```
vet.fit <- coxph(Surv(time, status) ~ trt, data = veteran)
summary(vet.fit)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ trt, data = veteran)
##
##   n= 135, number of events= 126
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## trt 0.1354    1.1450   0.1800 0.752    0.452
##
##     exp(coef) exp(-coef) lower .95 upper .95
## trt     1.145     0.8734    0.8047     1.629
##
## Concordance= 0.534  (se = 0.026 )
## Likelihood ratio test= 0.56  on 1 df,   p=0.5
## Wald test            = 0.57  on 1 df,   p=0.5
## Score (logrank) test = 0.57  on 1 df,   p=0.5
```

This output is essentially comparing treatment 2 (chemotherapy) to treatment 1 (standard). The hazard ratio given by the summary for "trt" is 1.1450; this comes from the exp(coef) part of the output. This tells us that treatment 2 (chemotherapy) has a higher risk of death than treatment 1 (standard). However, this summary also gives a p-value of 0.452. Thus, the difference between the survival rates of patients receiving the chemotherapy treatment and the standard treatment is statistically insignificant.

The 95% confidence interval for this hazard ratio is (0.8047, 1.629). 1 is included in this interval, which also suggests it is insignificant. This interval comes directly from the summary output.

## 2.4 Likelihood Ratio Test

Next, we will look at the log-likelihood for the model used in the previous section.

```
anova(vet.fit)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##       loglik Chisq Df Pr(>|Chi|)
## NULL -496.04
## trt  -495.76 0.565  1     0.4523
```

The log-likelihood is -495.76. It has a p-value of 0.4523 which is greater than 0.05. Therefore, we fail to reject the null hypothesis.

# 3 Model Building

Next, we will be model building by forward selecting for covariates and analyzing their AIC's. For every model, we will keep the "trt" covariate at the end of the equation to control for it. We do this because the treatment type is the important question that we are trying to solve for. In general, we are looking for a model with the lowest AIC value. If we add a new covariate and the AIC value increases or barely changes at all, then we will go back and choose the previous model.

| Model | df | AIC | covariates | chosen? |
|---|---|---|---|---|
| model1 | 4 | 977.2560 | celltype + trt | no |
| model2 | 2 | 958.3115 | karno + trt | yes |
| model3 | 2 | 994.0227 | diagtime + trt | no |
| model4 | 2 | 995.4603 | age + trt | no |
| model5 | 2 | 995.2560 | prior + trt | no |
| model2.1 | 5 | 948.6572 | karno + celltype + trt | yes |
| model2.3 | 3 | 960.2443 | karno + diagtime + trt | no |
| model2.4 | 3 | 959.7413 | karno + age + trt | no |
| model2.5 | 3 | 959.9257 | karno + prior + trt | no |
| model2.1.3 | 6 | 950.4752 | karno + celltype + diagtime + trt | no |
| model2.1.4 | 6 | 949.2749 | karno + celltype + age + trt | no |
| model2.1.5 | 6 | 949.2992 | karno + celltype + prior + trt | no |
| model2.1.int1 | 6 | 949.7646 | karno + celltype + trt:karno + trt | no |
| model2.1.int2 | 8 | 949.2949 | karno + celltype + trt:celltype + trt | no |

In the end, the best model that we found was model 2.1.

## 3.1 Analysis of Model

Now that we have decided on a model, we will look into how each covariate contributes to the overall model.

```
summary(model2.1)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ karno + celltype + trt,
##     data = veteran)
##
##   n= 135, number of events= 126
##
##                         coef exp(coef)  se(coef)      z Pr(>|z|)
## karno              -0.029322  0.971104  0.005201 -5.637 1.73e-08 ***
## celltypesmallcell   0.739184  2.094226  0.264490  2.795 0.005194 **
## celltypeadeno       1.041102  2.832336  0.292343  3.561 0.000369 ***
## celltypelarge       0.262685  1.300417  0.280137  0.938 0.348397
## trt                 0.321312  1.378936  0.199460  1.611 0.107199
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
##                    exp(coef) exp(-coef) lower .95 upper .95
## karno                 0.9711     1.0298    0.9613    0.9811
## celltypesmallcell     2.0942     0.4775    1.2471    3.5169
## celltypeadeno         2.8323     0.3531    1.5970    5.0233
## celltypelarge         1.3004     0.7690    0.7510    2.2518
## trt                   1.3789     0.7252    0.9327    2.0386
##
## Concordance= 0.733  (se = 0.02 )
## Likelihood ratio test= 53.43  on 5 df,   p=3e-10
## Wald test            = 55.65  on 5 df,   p=1e-10
## Score (logrank) test = 58.15  on 5 df,   p=3e-11
```

The p-value of the likelihood-ratio test, Wald test, and score test are significant, indicating that the overall model is significant in predicting the survival function. The p-value for karno, celltypesmallcell, and celltypeadeno are all less than 0.05, indicating they are significant. The p-value for celltypelarge and treatment are both greater than 0.05, so they fail to be significant.

The p-value for karno is 1.73e-08 with a hazard ratio of 0.9711, indicating that there is a strong relationship between a patient's Karnofsky performance score and decreased risk of death.

The p-value for celltypesmallcell is 0.0051984 with a hazard ratio of 2.094226, indicating that there is a strong relationship between a patient having the small cell type and increased risk of death.

The p-value for celltypeadeno is 0.000369 with a hazard ratio of 2.832336, indicating that there is a strong relationship between a patient having the adeno cell type and increased risk of death.

Unlike the other covariates, celltypelarge and trt have p-values of 0.348397 and 0.107199 and hazard ratios of 1.300417 and 1.378936, respectively. They have 95% confidence intervals of (0.7510, 2.2518) and (0.9327, 2.0386), respectively. Both of these intervals include 1, indicating that both celltypelarge and trt do not significantly contribute to a patient's risk of death.

To reiterate, from the summary of our model, we found that when we account for karno and celltype, **treatment is not a significant covariate.**

However, before we can be confident in the results provided above, we must first test the assumptions of the Cox PH model.

# 4 Assumptions of the Cox PH model

The main assumption of the Cox PH model that we will test for is the proportional hazards assumptions. In order to test for this, we will use log-log plots and the Schoenfeld's residuals test. If either of these tests suggest a violation in the proportional hazards assumptions, then we can solve this by stratifying on the problem variable or making a time dependent covariate. First, we will look at log-log plots for our variables.

## 4.1 Log-log Plot for Covariates

```
ll.plot <- survfit(Surv(time, status) ~ trt, data = veteran)

plot(ll.plot, fun = "cloglog",
     main = "Log-log Plot for Treatment", xlab = "log(time)", ylab = "log(-log(S(t)))")
```

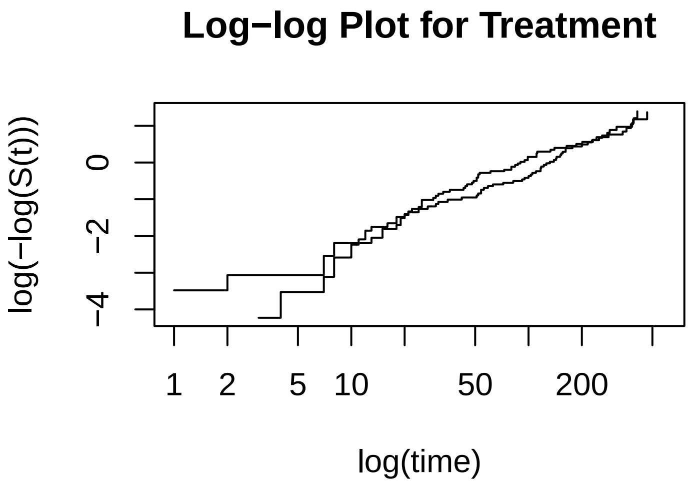


Figure 3: Log-log plot for 'trt' covariate.

```r
plot(survfit(Surv(time, status) ~ karno, data = veteran), fun = "cloglog",
     main = "Log-log Plot for Karno", xlab = "log(time)", ylab = "log(-log(S(t)))")
```

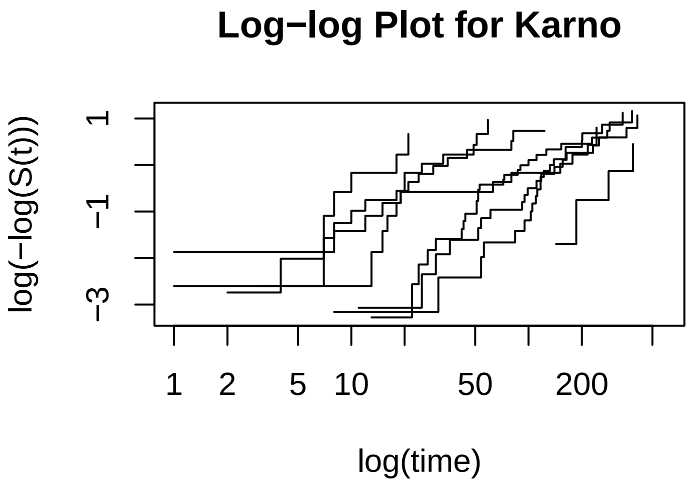


Figure 4: Log-log plot for 'karno' covariate.

```
plot(survfit(Surv(time, status) ~ celltype, data = veteran), fun = "cloglog",
     main = "Log-log Plot for Cell Type", xlab = "log(time)", ylab = "log(-log(S(t)))")
```

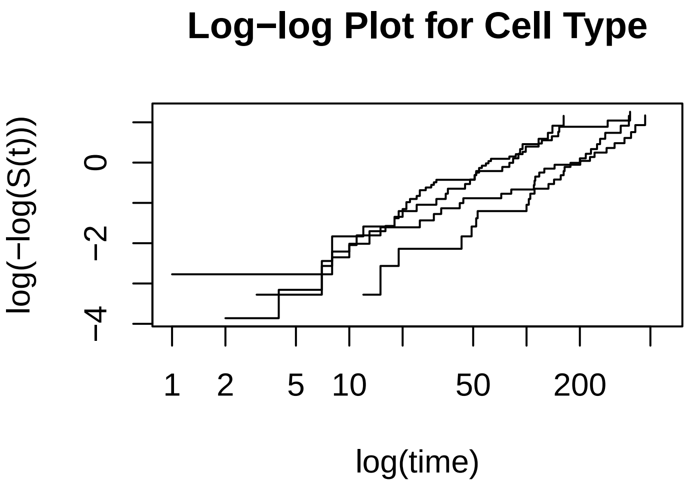


Figure 5: Log-log plot for ‘celltype’ covariate.

In general, the log-log plot tells us that if all of these lines are parallel, then the Cox PH assumption is appropriate.

The log-log plot for treatment (Figure 3) appears to be parallel and thus does not violate the Cox PH assumption. However, the log-log plots for karno and celltype (Figures 4 and 5) are questionable, so it seems to violate the proportional hazards assumption.

We will check this further by looking at the Schoenfeld residuals.

### 4.2 Schoenfeld residuals

```
sch.res <- cox.zph(model2.1)
sch.res
```

```
##           chisq df       p
## karno    17.49  1 2.9e-05
## celltype 12.68  3 0.00538
## trt       0.57  1 0.45020
## GLOBAL   24.38  5 0.00018
```

```
plot(sch.res[3], main = "Schoenfeld's Residuals")
abline(h = 0, col = "blue")
```

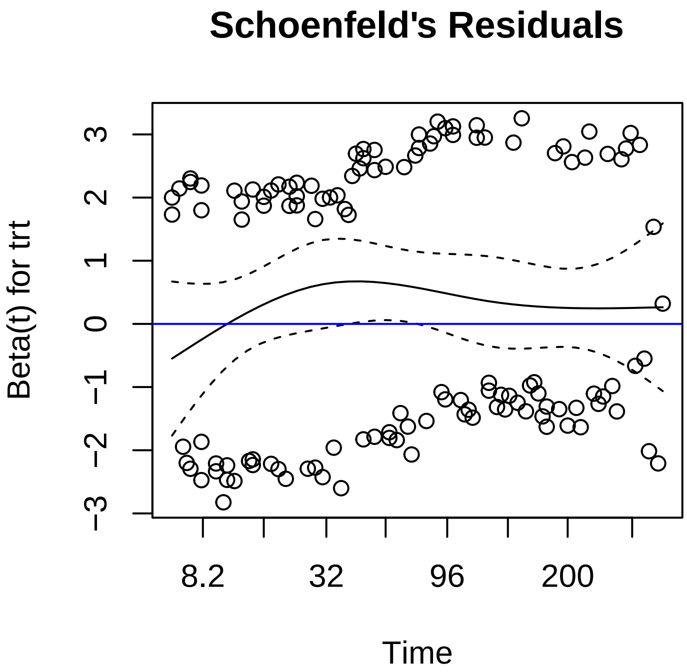


Figure 6: Checking Schoenfeld's residuals for 'trt' covariate.

```
plot(sch.res[2], main = "Schoenfeld's Residuals")
abline(h = 0, col = "blue")
```

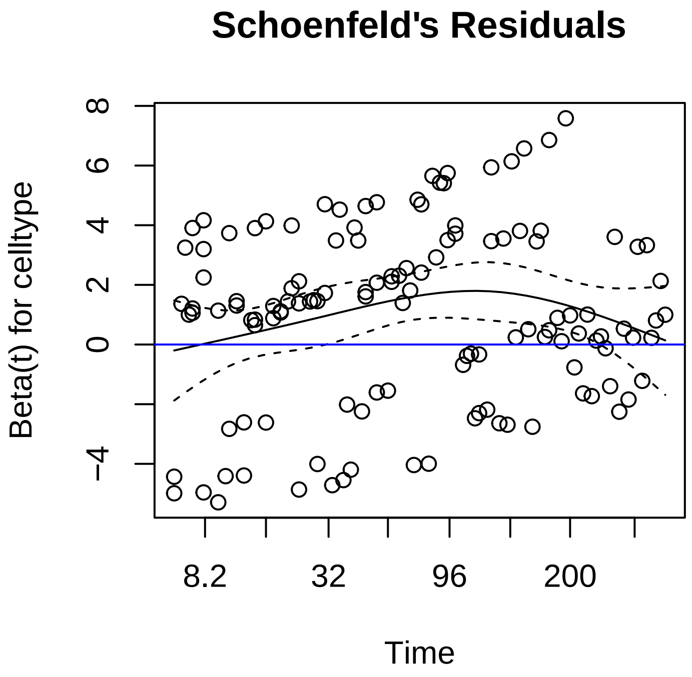


Figure 7: Checking Schoenfeld's residuals for 'celltype' covariate.

```
plot(sch.res[1], main = "Schoenfeld's Residuals")
abline(h = 0, col = "blue")
```

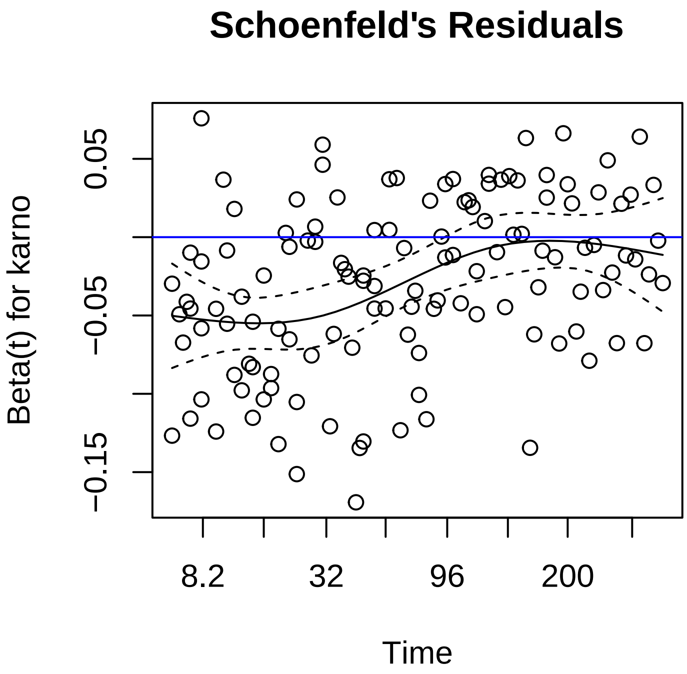


Figure 8: Checking Schoenfeld's residuals for 'karno' covariate.

The p-value for treatment (**Figure 6**) is 0.4502, which is greater than 0.05. The plot also appears to show no pattern with time. Thus, treatment appears to support the PH assumption.

However, both karno and celltype (**Figures 7 and 8**) have p-values much lower than 0.05. This indicates that the residuals for these covariates change with time and as a result, the PH assumption is likely violated.

### 4.3 Resolving Assumptions

In order to resolve these violations of the proportional hazards assumption, we can either add a covariate*time interaction, or we can stratify on the covariates. Because both karno and celltype are not the covariates that we are directly interested in, we will opt to stratify on them. This will cause us to lose some information, but it will allow us to use the PH assumption.

Stratifying on problem covariates:

```
strat.model2.1 <- coxph(Surv(time, status) ~ strata(karno) + strata(celltype) + trt, data = veteran)
summary(strat.model2.1)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ strata(karno) + strata(celltype) +
##     trt, data = veteran)
##
##   n= 135, number of events= 126
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## trt 0.1872    1.2059   0.2626 0.713    0.476
```

```
##
##     exp(coef) exp(-coef) lower .95 upper .95
## trt     1.206     0.8292    0.7207     2.018
##
## Concordance= 0.524  (se = 0.037 )
## Likelihood ratio test= 0.51  on 1 df,   p=0.5
## Wald test            = 0.51  on 1 df,   p=0.5
## Score (logrank) test = 0.51  on 1 df,   p=0.5
```

The above code tells us that after stratifying on karno and celltype, treatment still appears to not be a significant covariate. It has a p-value of 0.476 and a hazard ratio of 1.2059 with a 95% confidence interval of (0.7207, 2.018). This interval includes 1, indicating that the treatment does not significantly contribute to a patient's risk of death.

Quick assumption check:

```
sch.res2 <- cox.zph(strat.model2.1)
sch.res2
```

```
##        chisq df    p
## trt    0.657  1 0.42
## GLOBAL 0.657  1 0.42
```

The quick check on the Schoenfeld residuals looks good.

# 5 Advanced Methods

So far throughout this report, we have found nothing to indicate that the chemotherapy treatment is significantly more effective at lowering patient's risk compared to the standard treatment. However, in order to investigate this further, we will look to see if treatment is a time-varying covariate.

We will split our data at about 3 months and at about 6 months to see if the treatment is more effective during any of these time periods: (0,90], (90, 180], (180, . . . ).

```
veteran.split <- survSplit(Surv(time, status) ~ ., data= veteran, cut=c(90, 180),
episode= "tsplit", id="id")

vet.splitfit <- coxph(Surv(tstart, time, status) ~ strata(karno) + strata(celltype) +
                        trt*strata(tsplit),
               data = veteran.split)
vet.splitfit
```

```
## Call:
## coxph(formula = Surv(tstart, time, status) ~ strata(karno) +
##     strata(celltype) + trt * strata(tsplit), data = veteran.split)
##
##                                 coef exp(coef) se(coef)      z      p
## trt                           0.4712    1.6019   0.3151  1.495 0.1348
## trt:strata(tsplit)tsplit=2 -1.2991    0.2728   0.7498 -1.732 0.0832
## trt:strata(tsplit)tsplit=3 -0.4712    0.6243   0.8897 -0.530 0.5964
##
## Likelihood ratio test=3.91  on 3 df, p=0.2715
## n= 219, number of events= 126
```

After splitting the data and looking at the Cox PH model, we can see that the p-values given are 0.1348, 0.0832, 0.5964 for the three respective time periods. Therefore, for each time period – 0-90 days, 90-180 days, and 180 days and on – treatment remains an insignificant covariate. Again, because we stratified on karno and celltype, we are unable to say anything about them.

```r
anova(vet.splitfit)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(tstart, time, status)
## Terms added sequentially (first to last)
##
##                       loglik  Chisq Df Pr(>|Chi|)
## NULL                 -121.56
## trt                  -121.31 0.5079  1     0.4760
## trt:strata(tsplit) -119.61 3.4006  2     0.1826
```

Checking the log-likelihood briefly shows that both "trt" and "trt:strata(tsplit)" have p-values greater than 0.05. This looks good.

Next, we will check Schoenfeld residuals.

```r
cox.zph(vet.splitfit)
```

```
##                      chisq df    p
## trt                  0.762  1 0.38
## trt:strata(tsplit) 1.621  2 0.44
## GLOBAL               2.045  3 0.56
```

Quickly checking the Schoenfeld residuals, we see that the p-value for "trt" and "trt:strata(tsplit)" are greater than 0.05, meaning that our proportional hazards assumption holds.

Next, we will attempt a split at every time where there is an event. This method is a lot more "expensive" in terms of computing power, but it will give us another view to look at the significance (or lack thereof) of our treatment covariate.

```r
all.times <- veteran$time[veteran$status == 1]

vet.bigsplit <- survSplit(Surv(time, status) ~ .,
                          data = veteran,
                          cut = all.times,
                          id = "Subject",
                          episode = "tsplit")

vet.bigfit <- coxph(Surv(tstart, time, status) ~ strata(karno) + strata(celltype) + trt*tsplit,
                    data = vet.bigsplit)
vet.bigfit
```

```
## Call:
## coxph(formula = Surv(tstart, time, status) ~ strata(karno) +
##     strata(celltype) + trt * tsplit, data = vet.bigsplit)
##
##                   coef  exp(coef)  se(coef)       z     p
## trt          0.479901   1.615914   0.449536   1.068 0.286
## tsplit            NA        NA  0.000000     NA    NA
## trt:tsplit -0.008818  0.991221  0.010978 -0.803 0.422
##
## Likelihood ratio test=1.16  on 2 df, p=0.5604
## n= 5766, number of events= 126
```

After splitting the time at every event point, our treatment*tsplit covariate resulted in a hazard ratio of 0.991221 and a p-value of 0.422. Again, we see that the p-value is greater than 0.05. Therefore, there is no significant difference between the chemotherapy treatment and the standard treatment.

# 6 Conclusion

The goal of this report was to investigate whether or not chemotherapy treatment (trt 2) was significantly better than the standard treatment (trt 1) when it comes to treating lung cancer.

In our best model, we used the "karno", "celltype", and "trt" covariates. However, "karno" and "celltype" violated our proportional hazards assumptions, so we decided to stratify on both those covariates. In our stratified model, the treatment covariate had a p-value of 0.476 and a hazard ratio of 1.2059 with a 95% confidence interval of (0.7207, 2.018). This interval includes 1, indicating that the treatment does not significantly contribute to a patient's risk of death. Despite attempting a couple different time-varying covariate models, the p-value for treatment never dropped below 0.05.

Ultimately, through all of our models and tests, we discovered that treatment was not a significant covariate. Thus, there is **no significant difference between the chemotherapy treatment and the standard treatment.**

# 7 References

Therneau, T., Crowson, C., & Atkinson, E. (2022, August 5). Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model.

D Kalbfleisch and RL Prentice (1980), The Statistical Analysis of Failure Time Data. Wiley, New York.