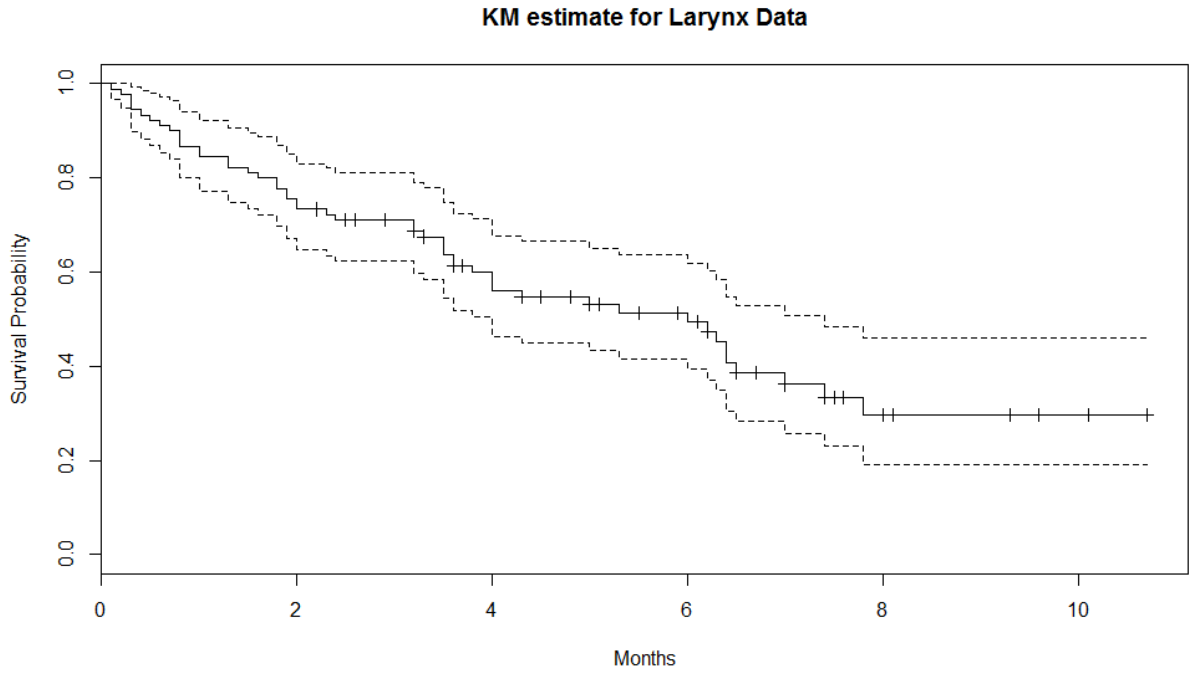
PSTAT 175 Final Project

Introduction

Our project is using data from the package in the "KMsurv" package: larynx. In this data set, there are 90 observations for larynx cancer patients at different ages, stages, and year diagnosed in the 1970s, as well as their survival time (in months) and a death indicator.

To begin, as a whole, the Kaplan-Meier estimate illustrates the data's survival estimates by their times and death indicator:

```
> plot(survfit(Surv(time,delta)~1,data=larynx), xlab="Months", ylab="Survival Probability", main="KM estimate for Larynx Data")
```



This plot KM estimator can be used to estimate many different, basic calculations.
This plot can be useful in finding the estimate for the survivor function at 3.0 months.

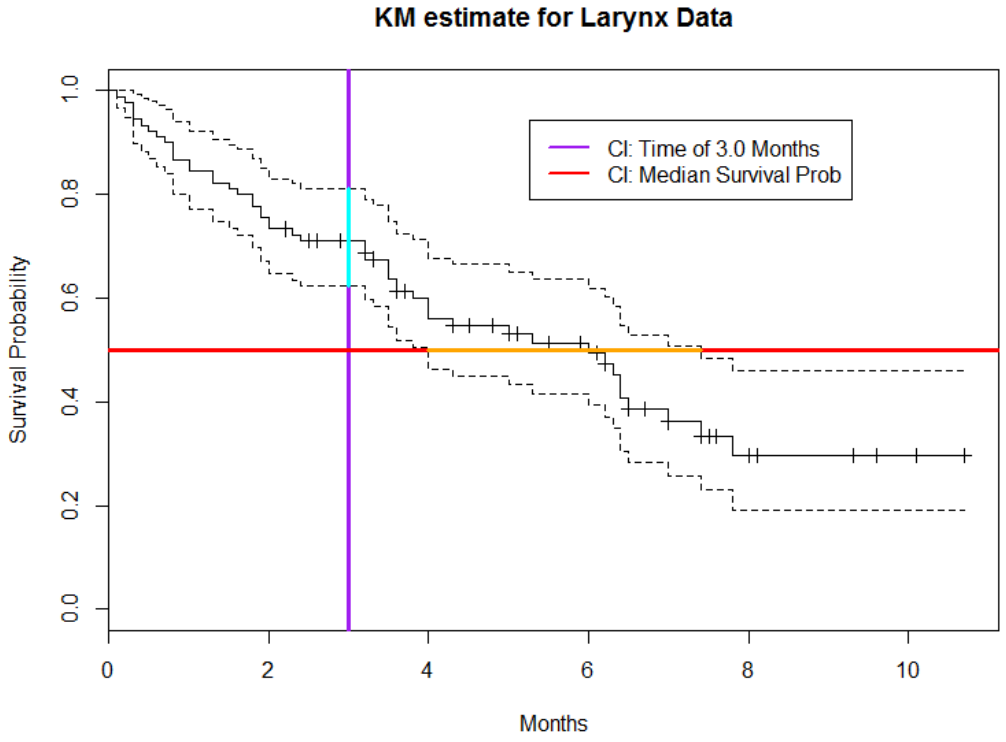
```
> summary(survfit(Surv(time,delta)~1, data=larynx),time=3.0)
Call: survfit(formula = Surv(time, delta) ~ 1, data = larynx)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
3	60	26	0.711	0.0478	0.623	0.811

The estimated survivor function at time 3 months = 0.711. The 95 % confidence interval for the survivor function at 3 months: (0.623, 0.811)

```
> plot(lFit, xlab="Months", ylab="Survival Probability", main="KM estimate for Larynx Data")
> summary(survfit(Surv(time,delta)~1, data=larynx),time=3.0)
```

```
> abline(v=3,lwd=3,col="purple")
> segments(3,.623,3,.811,lwd=3,col=5)
> abline(h=0.5,lwd=3,col="red")
> segments(4,.5,7.4,.5,lwd=3,col="orange")
> legend(locator(1),legend=c("CI: Time of 3.0 Months","CI: Median Survival Prob"), lwd=2, col=c("purple","red"))
```



A 95% confidence interval for the median of the data can be calculated. By looking at the summary(lFit), when t=6.0, this is the first step down that crosses below 0.50. For the confidence interval, by looking at the lower and upper bounds for the times, all of the intervals containing 0.50 are when times are between 4.0 and 7.4. Thus the 95 %CI for the median is [4.0,7.4].

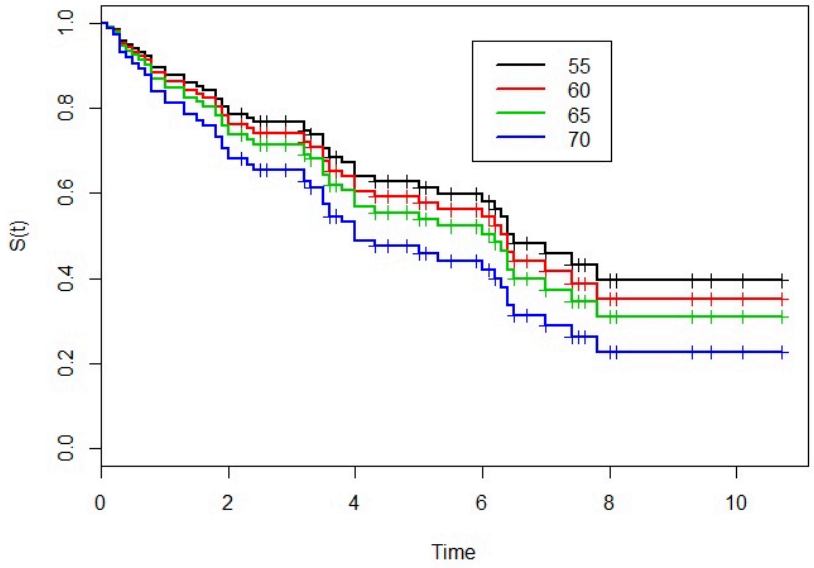
There are many different applications and calculations of the data can be used to illustrate a concept or to obtain a greater understanding about the circumstances.

There are 3 covariates included in the larynx data set: stage, age, and year diagnosed. In this project we will determine which covariates play a significant role in the model.

Model Selection

First we wanted to determine if age has a significant effect on the survival time.

```
> fit.a<-coxph(Surv(time,delta)~age, data=larynx)
> plot(survfit(fit.a, data.frame(age=c(55,60,65,75))),lwd=2, col=c(1,2,3,4),xlab="Time", ylab="S(t)")
> legend(locator(1), legend=c("55","60","65","70"),col=c(1,2,3,4),lwd=2)
```

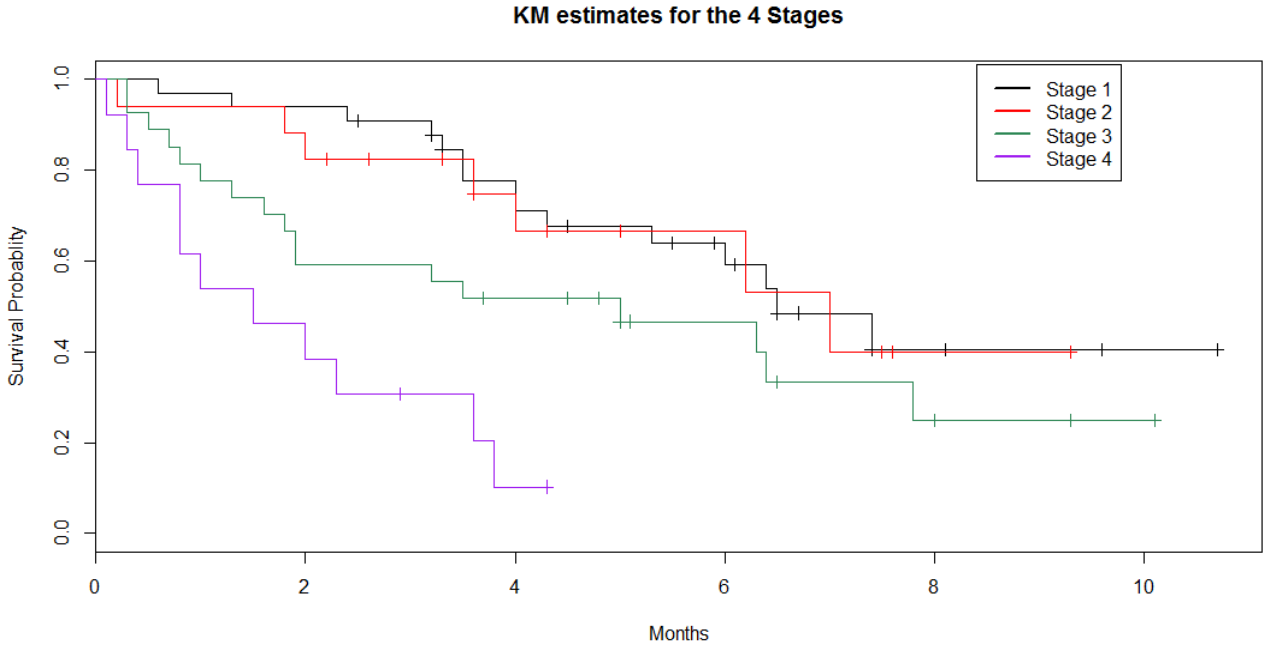



```
> fit.a
Call:
coxph(formula = Surv(time, delta) ~ age, data = larynx)
          coef  exp(coef)   se(coef)     z        p
age     0.0233       1.02     0.0145  1.61     0.11
Likelihood ratio test=2.63  on 1 df, p=0.105  n= 90, number of events= 50
```

In the hypothesis test with respect to the beta value for age, the null hypothesis is that beta for age is 0, meaning age has no effect in the model. The alternative hypothesis is that the beta value for age does not equal to 0, meaning age has a significant effect in the model.
By the likelihood ratio test, the statistic is 2.63, its p-value is 0.1048. Since the p-value is greater than alpha, then the null hypothesis is concluded: age has no effect on the model.

We then wanted to see if stage played a significant role in the model.

```
> plot(lFit, col=c("black", "red", "seagreen", "purple"), lwd=c(1,1,1,1), xlab= "Months", ylab="Survival Probability",
main="KM estimates for the 4 Stages")
> legend(locator(1), legend=c("Stage 1", "Stage 2", "Stage 3", "Stage 4"), col=c("black", "red", "seagreen", "purple"),
lwd=c(2,2,2,2))
```




```
> fit.s
Call:
coxph(formula = Surv(time, delta) ~ stage, data = larynx)
            coef  exp(coef)   se(coef)    z       p
stage      0.509     1.66        0.141     3.6     0.00031
Likelihood ratio test=13.3  on 1 df, p=0.000271  n= 90, number of events= 50
```

Null hypothesis: beta for stage is 0, alternative hypothesis: beta for stage is not 0.
Since the p-value for this test is 0.000271, this value is less than alpha. Thus the alternative hypothesis is concluded: beta for stage is not 0, thus stage has a significant effect in the model

Lastly for the covariate year diagnosed,

```
> fit.d
Call:
coxph(formula = Surv(time, delta) ~ diagyr, data = larynx)
            coef      exp(coef)      se(coef)      z          p
diagyr     0.022         1.02          0.0717     0.307     0.76
Likelihood ratio test=0.09  on 1 df, p=0.759  n= 90, number of events= 50
```

Null hypothesis: beta for diagnosis year is 0, alternative hypothesis: beta for diagnosis year is not 0.Since the p-value for this test is 0.759, this value is greater than alpha. Thus the null hypothesis is concluded: beta for diagnosis year is 0, thus the year of diagnosis has no effect in the model.

But testing the covariates doesn't necessarily mean they are to be added into the model or not. So an additional method is needed to determine which should be included in the model.

To find the best, but simplest, model of fit for the data (and for any other data set with the same manner) is to be determined by either forward or backward selection. For small models, as larynx is, the best method for model selection is forward selection: start with no covariates, add one at a time where the best is added first, and then test whether to stop adding covariates in the model or to continue the selection.

First step is to find the likelihood value for each covariate.

> fit.a<-coxph(Surv(time, delta)~age, data=larynx) > fit.d<-coxph(Surv(time, delta)~diagyr,data=larynx) > fit.s<-coxph(Surv(time, delta)~stage,data=larynx)	> fit.a$loglik[2] [1] -195.5478 > fit.d$loglik[2] [1] -196.8163 > fit.s$loglik[2] [1] -190.2317

The best likelihood of the three covariates is the one with the largest (least negative) value: our result is stage with loglik= -190.2317. The test used to determine whether to add stage or not is the LRT for stage. The null hypothesis is all 3 beta values for stage, age, and year diagnosed are 0, meaning none of these covariates have any significant effect in the model and to look at other variables. The alternative hypothesis is the beta value for stage is not 0 while the beta values for age and year diagnosed are 0. Since the LRT (from fit.a) is 13.26, the p-value is 0.0002706. Since the p-value is less than alpha, the alternative hypothesis is concluded: the beta value for stage is not 0 and the beta values for age and year diagnosed are 0.

Since the p-value is less than alpha, forward selection is continued.

The second step is to find the next best covariate.

> fit.s.a<-coxph(Surv(time,delta)~stage+age, data=larynx) > fit.s.d<-coxph(Surv(time,delta)~stage+diagyr,data=larynx)	> fit.s.a$loglik[2] [1] -188.9517 > fit.s.d$loglik[2] [1] -190.2128

The best likelihood in this step is age with a likelihood value of -188.9517. Next is to test whether to include age in the model or to not include age and stop the selection. The null hypothesis for this test is that only the beta value for stage is not equal to 0, while the alternative hypothesis is that only stage and age have betas not equal to 0.
Null hypothesis: fit.s$loglik[2], and alternative hypothesis: fit.s.a$loglik[2]

```
> test<- 2*(fit.s.a$loglik[2] - fit.s$loglik[2])
> test
[1] 2.559811
> pchisq(test, df = 2-1, lower.tail=FALSE)
[1] 0.1096117
```

Since the p-value is greater than alpha, age is not added into the model and model selection is stopped. The best fit and simplest model for of the larynx data set, and any other data set associated, is with the covariate of stage of the disease.

Stratified Models

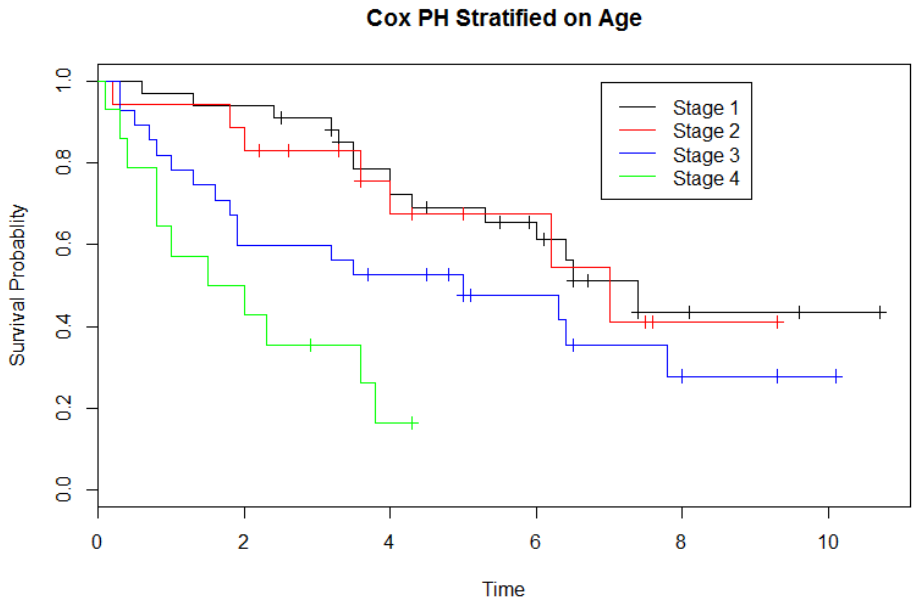
Lastly, we want to test if splitting the data into subsets based on stages makes the covariate age significant. By doing this, we have a different baseline for each age, but still only get one parameter, beta. The partial likelihood is calculated separately for each stage and then estimates beta from the partials.

```
> strat.fit.age<-coxph(Surv(time, delta)~strata(stage) + age ,data=larynx)

Call:
coxph(formula = Surv(time, delta) ~ strata(stage) + age, data = larynx)
          coef        exp(coef)      se(coef)       z          p
age     0.0166          1.02           0.0142       1.17     0.24
Likelihood ratio test=1.39  on 1 df, p=0.238  n= 90, number of events= 50
```

The null hypothesis: beta for age is equal to 0, which means that age has no effect on the model. The alternative hypothesis: beta for age does not equal 0, which means that age does have a significant effect on the model.
Since the p-value for the Likelihood Ratio Test is 0.238, we accept the null hypothesis, which means that age does not have a significant effect on the model when stratified for stages. We still see a significant effect of stage on the model, but we now know that stage plays the significant role, while age does not.

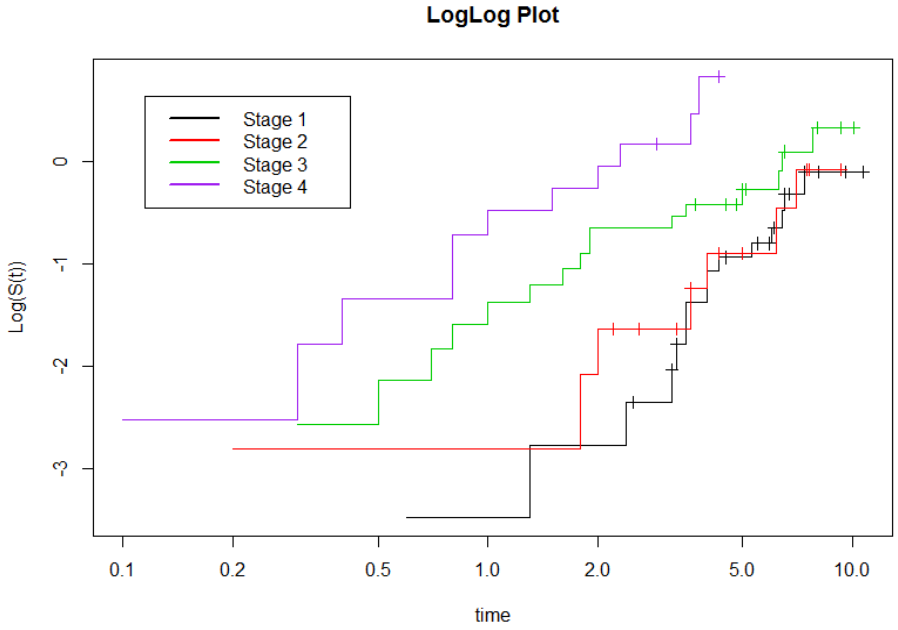
```
> plot(survfit(strat.fit.age), col= c("black", "red", "blue", "green"), lwd=c(1,1,1,1), xlab= "Time", ylab="Survival
Probability", main="Cox PH Stratified on Age")
> legend(locator(1), legend=c("Stage 1", "Stage 2", "Stage 3", "Stage 4"), col=c("black", "red", "blue", "green"),
lwd=c(1,1,1,1))
```



Model Checking

Though not as significant, it is good to check the model and check the goodness of fit.
This checking can be done in 2 ways.
The first method is to plot the model by plotting the 4 stages in a log-log plot.

```
> plot(survfit(Surv(time,delta)~stage,data=larynx), fun="cloglog", xlab="time", ylab="Log(S(t))", main="LogLog Plot",
col=c(1,2,3,"purple"))
> legend(locator(1),legend=c("Stage 1", "Stage 2", "Stage 3", "Stage 4"), col=c(1,2,3,"purple"),lwd=2)
```



Because stages 1 and 2 cross and overlap in many areas, it is wise to combine the 2 stages.

Another method of checking the model is to test using cox.zph. The null hypothesis for this test is the model is fine, the coxph is correct, and the proportional hazards assumption is correct. The alternative hypothesis is the proportional hazards assumption is rejected and the model is not good.

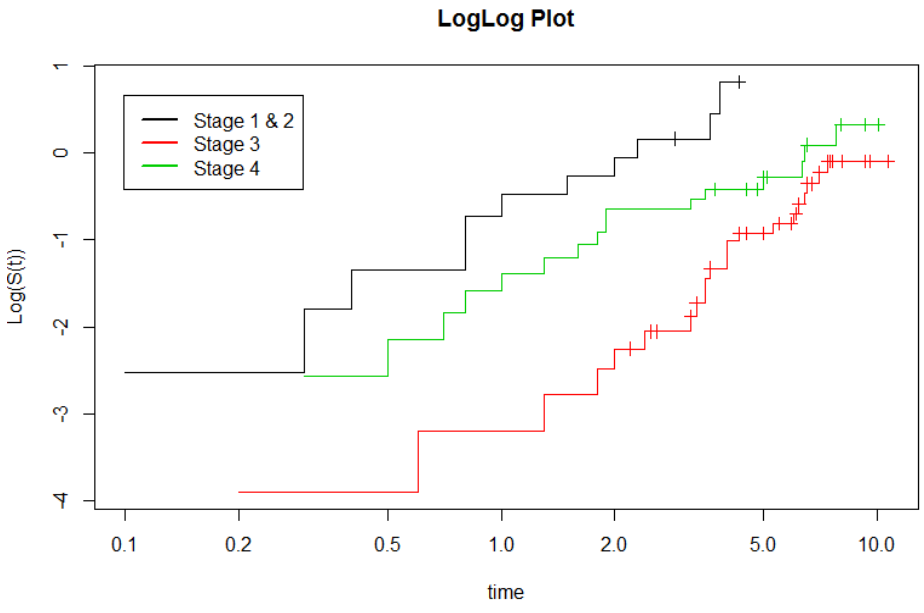
```
> cox.zph(fit.s)
        rho   chisq       p
stage -0.288    4.49  0.0342
```

Since the p-value for this test is 0.0342, the p-value is less than alpha. Thus the alternative hypothesis is concluded and this model is not good.
This results in combining Stage 1 and Stage 2, which was previously stated because their log log lines cross.

Altering Stage Levels

```
> larynx$stage2<-larynx$stage
> larynx$stage2[larynx$stage=="1"]<-"2"
> levels(larynx$stage2)<-c("1+2", "3","4")
> coxph(Surv(time,delta)~stage2,data=larynx)
Call:
coxph(formula = Surv(time, delta) ~ stage2, data = larynx)
             coef  exp(coef)   se(coef)     z        p
stage23    0.595       1.81      0.324  1.84  6.6e-02
stage24    1.714       5.55      0.391  4.38  1.2e-05
Likelihood ratio test=16.5  on 2 df, p=0.000266  n= 90, number of events= 50
```

```
> plot(survfit(Surv(time,delta)~stage2,data=larynx), fun="cloglog", xlab="time", ylab="Log(S(t))", main="LogLog Plot",
col=c(1,2,3))
> legend(locator(1),legend=c("Stage 1 & 2", "Stage 3", "Stage 4"), col=c(1,2,3),lwd=2)
```



After factoring stage2 [> larynx$stage2<-factor(larynx$stage2)], when you plot the new log log plot of the 3 stages of the disease, the plot shows no crossings and lines parallel to each other. This is a good indicator that the proportional hazards model is a good fit.

Testing the New Stages

```
> fit.s2<-coxph(Surv(time,delta)~stage2,data=larynx)
> cox.zph(fit.s2)
                rho        chisq       p
stage23     -0.2712      3.519     0.0607
stage24     -0.0966      0.424     0.5152
GLOBAL         NA          3.521     0.1720
```

Since the p-value for Stage 3 compared to Stages 1 & 2 is barely greater than alpha, changing the baseline can strengthen the assumption.

To change the baseline to stage 4:

```
> larynx$stage2<-relevel(larynx$stage2,ref="4")
> new.sfit<-coxph(Surv(time,delta)~stage2,data=larynx)
> new.sfit
Call:
coxph(formula = Surv(time, delta) ~ stage2, data = larynx)
              coef      exp(coef)       se(coef)       z            p
stage22    -1.71        0.180            0.391       -4.38     1.2e-05
stage23    -1.12        0.326            0.406       -2.76     5.8e-03
Likelihood ratio test=16.5  on 2 df, p=0.000266  n= 90, number of events= 50
> cox.zph(new.sfit)
                rho          chisq          p
stage22      0.0966       0.424         0.515
stage23     -0.1198       0.754         0.385
GLOBAL         NA           3.521         0.172
```

Since the p-values for this new test are all greater than alpha, the null hypothesis for this test is concluded: this is a good model.

By the log log plot and the cox.zph test, this new model is a good fit model that holds the proportional hazards assumption true.

Conclusion

From our data, original graphs, and common sense, we knew that stage was going to play a significant role in our model. However, we wanted to see if age and the year the patient was diagnosed also played a crucial role in combination with stage.

An initial model selection was tested to see if age, stage, and year diagnosed play significant roles in the model. From these individual tests, stage was the only covariate that played a critical role.

After this, we did forward selection to build our model up from nothing. As suspected, stage was the first covariate added to our model. When age was tested (the next covariate that should have been added into the model), it was concluded that it was not significant to the model. Thus age was not added, and the selection process was stopped. From forward selection, we ended up with a model that included only the stage covariate.

Then we checked to see if model stratification would cause any covariate to be significant. We stratified on stage and tested for age. This still gave us the conclusion that age was not significant.

Then we checked this model to see if the proportional hazards assumption is correct. We determined that the proportional hazards model was not good if we used all 4 stages, but if we combined stages 1 and 2 together, the proportional hazards assumption was correct.

Every test we ran on the larynx data set provided us with the same conclusion: the only significant covariate is the stage of cancer. This is consistent with our prior beliefs that stage would play a critical role because people are placed into these categories based on the severity of their disease: the greater the severity, the less chance of survival.