Fall 2018

Survival Analysis Project

For

Diabetic Retinopathy

PSTAT 175

Drew Pritchard
Janelle Samansky
Issam Zejli

Background:

Diabetic Retinopathy is a disease that affects some people with diabetes. It occurs when high blood sugar damages the blood vessels in the retina. These blood vessels can swell and leak. Or they can close, stopping blood from passing through. All of these changes can cause loss of vision

Description:

This dataset was a 50 percent random sample of patients with "high-risk" diabetic retinopathy. The data contains 197 patients, and each one had two types of a laser treatments on a random eye (left eye or right eye). The event of interest was the time from initiation of treatment to the time when visual acuity dropped below 5/200. Censoring was caused by death, dropout, or end of the study.

Some of the scientific questions to answer in our analysis.

Is there a significant difference between the two laser treatments overall?
Is treatment at young age more effective than at an older age?
Is there a difference in treatments between left eye and right eye?
Are there any risk factors that dramatically influence the survival function('type', 'risk', etc.)

Complication

Each observation is actually one of two parts. We can see that there are two of each patient id. For each patient only one eye was treated(trt == 1). Essentially we are given a control for each patient. Generally a dataset will only include one observation per subject. This allows us to look for a discrepancy amongst treated eye and untreated eye survival times. This is especially convenient because there are no other variables to interfere with our inference as each patient has one treated and one untreated.

This does mean however that there is room for errors when performing our analysis and we have to be careful about the fact that our original dataset has a duplicate for each patient.

Dataset:

Retinopathy is a survival data for diabetic retinopathy, contains information such that

```
names(retinopathy)
```

```
## [1] "id"     "laser"  "eye"    "age"    "type"   "trt"
## [7] "futime" "status" "risk"
```

```
head(retinopathy)
##    id laser   eye    age    type  trt futime     status     risk
## 1  5 argon   left    28          adult   1  46.23       0        9
## 2  5 argon   left    28          adult   0  46.23       0        9
## 3 14 argon  right  12        juvenile   1  42.50       0        8
## 4 14 argon  right  12        juvenile    0  31.30       1        6
## 5 16 xenon right   9        juvenile   1  42.27       0       11
## 6 16 xenon right   9        juvenile    0  42.27       0       11
```

Description of variables:

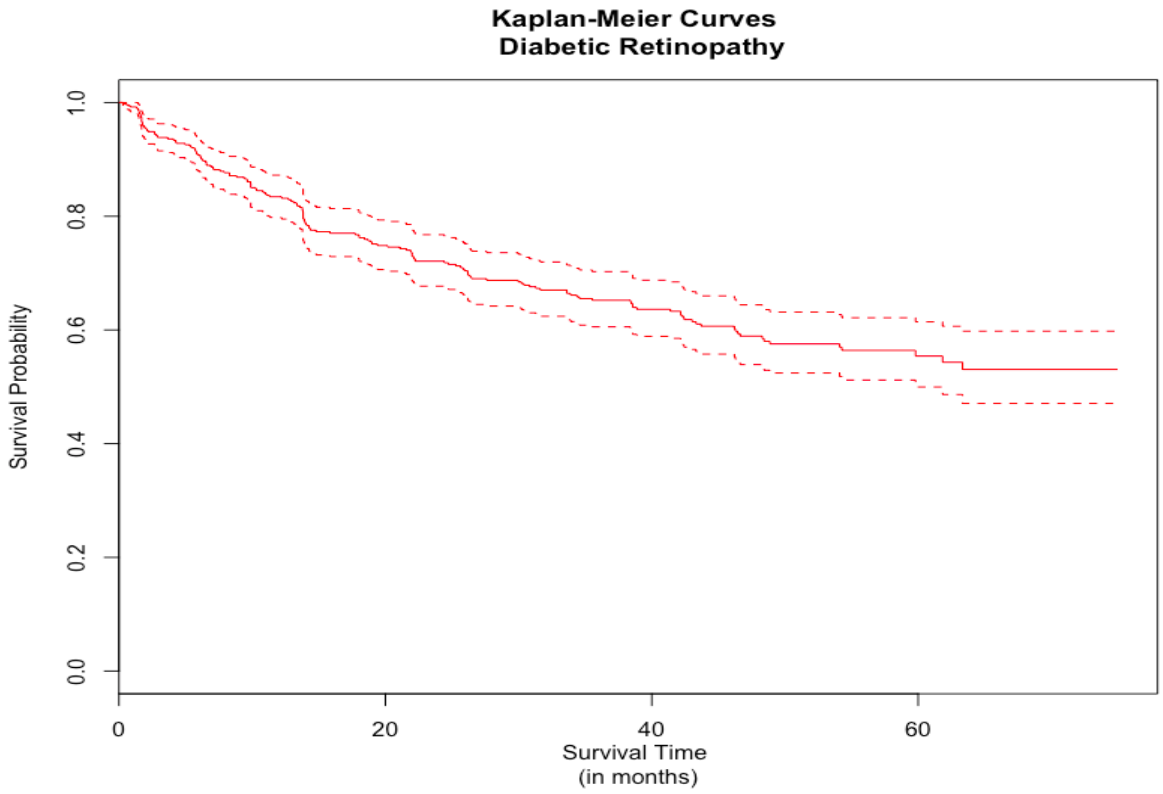
Laser: There are two type of laser used for treatments: argon and xenon.
Eye: which eye is treated left or right.
Type: type of diabetes (juvenile adult)
Age: age at diagnosis of diabetes
Trt: 0 = control eye, 1 = treated eye
Futime = Time to loss of vision or last follow-up
Status : 0 = censored, 1 = loss of vision in this eye
Risk = risk score for the eye. (This high-risk subset is defined as a score of 6 or greater in at least one eye.)

General analysis

KM estimate: We use the Kaplan Meier curve to estimate the survival probability of Diabetic Retinopathy patients. As shown in the graph, the survival probability relatively decreases with each passing month.
Analyzing the quantile survival :the 10^{th} quantile for the survival function is the first time the survival function is below the probability of 0.9 which is equal to time 10.6.similarly for the 20^{th} quantile it is equal to time 13.38. In our data, no observation whose event occurred at time 61 However, KM method assumes that the survival time is the same as the nearest point and in this case probability remain approximately 0.554

```
> quantile(ret.survfit, c(.1,.2,.3,.4,.5))
   10    20    30    40    50
 6.20 13.83 26.20 46.27    NA
```

Our study does not converge to a survival rate of 0 and about half of our patients survived until the end of the study. In this case we see that 90% make it past 6.2 months, 80% past 13.83, 70% past 26.20, 60% past 46.27, and after that we don't have anymore data.



Kaplan-Meier Analysis

The test statistic for the log-rank test is a large chi-square test that uses as its criterion a statistic that provides an overall comparison of the KM curves. The test compares the expected number of events with the actual number of events. The test won't be rejected as long as the expected number of events is the same as the actual number of events**.**

Our null hypothesis of the log-rank test is: H0: S1(t) = S2(t) (or h1(t) = h2(t))

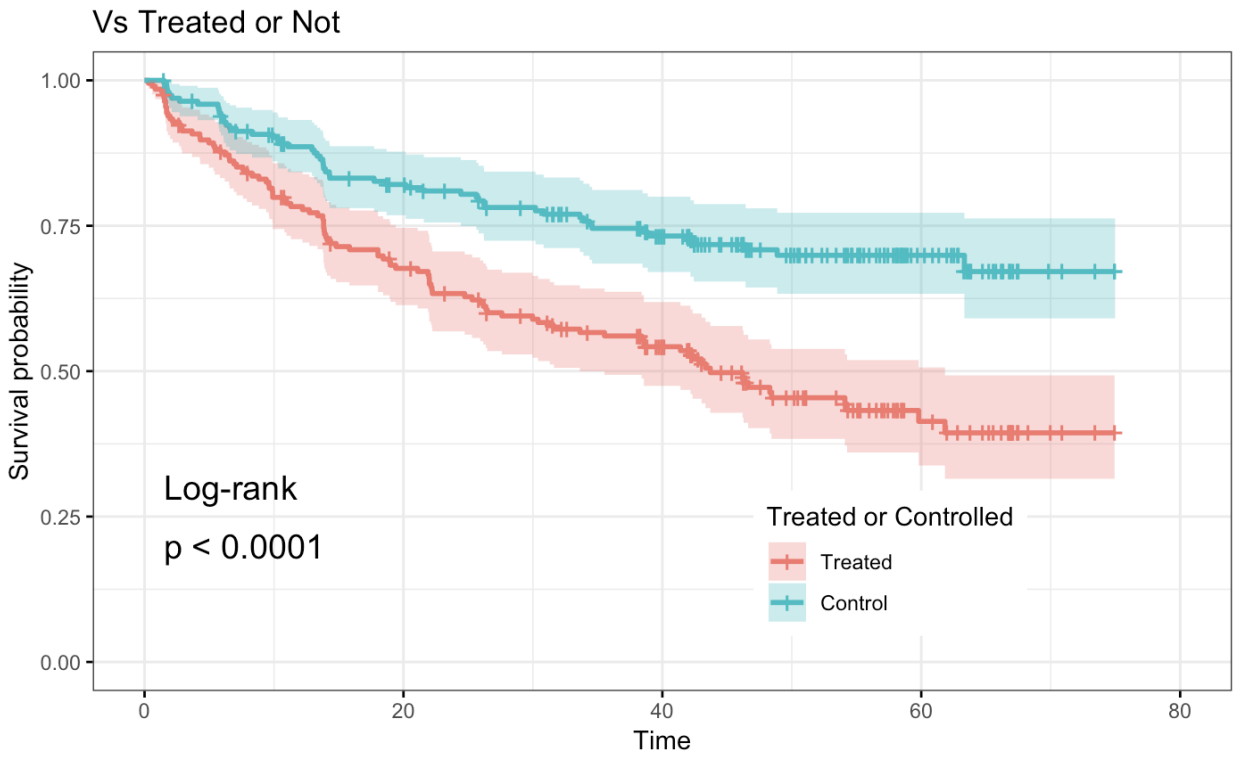
- The p value of the log- rank test for **laser** is 0.32 wish is greater than α = 5%, so we fail to reject the null hypothesis. Therefore, the survival function for laser Aaron is the same as the as the survival function for laser Xenon. Not significant difference.

-The p value of the log- rank test for **age** is 0.89 is also greater than α = 5%, so we fail to reject the null hypothesis. therefore, The survival function for age is the same, Not a significant difference.

-The p-value of the log- rank test for **Eye** is 0.09 wish is greater than α = 5%, so we fail to reject the null hypothesis. Therefore, the survival function for right eye is the same as the as the left eye. Not significant difference

-The p- value of log-rank test for trt and type is less than α = 5%, we fail to reject the null hypothesis, and conclude the survival functions are different and significant.

In order to visualize the results better, we use the function ggsurvplot in the survminer package. Looking at both graph 1 we can see that and 2 we can see the survival probability of the treated and untreated eye are different and the risk that risk 10 has the lowest survival probabilty



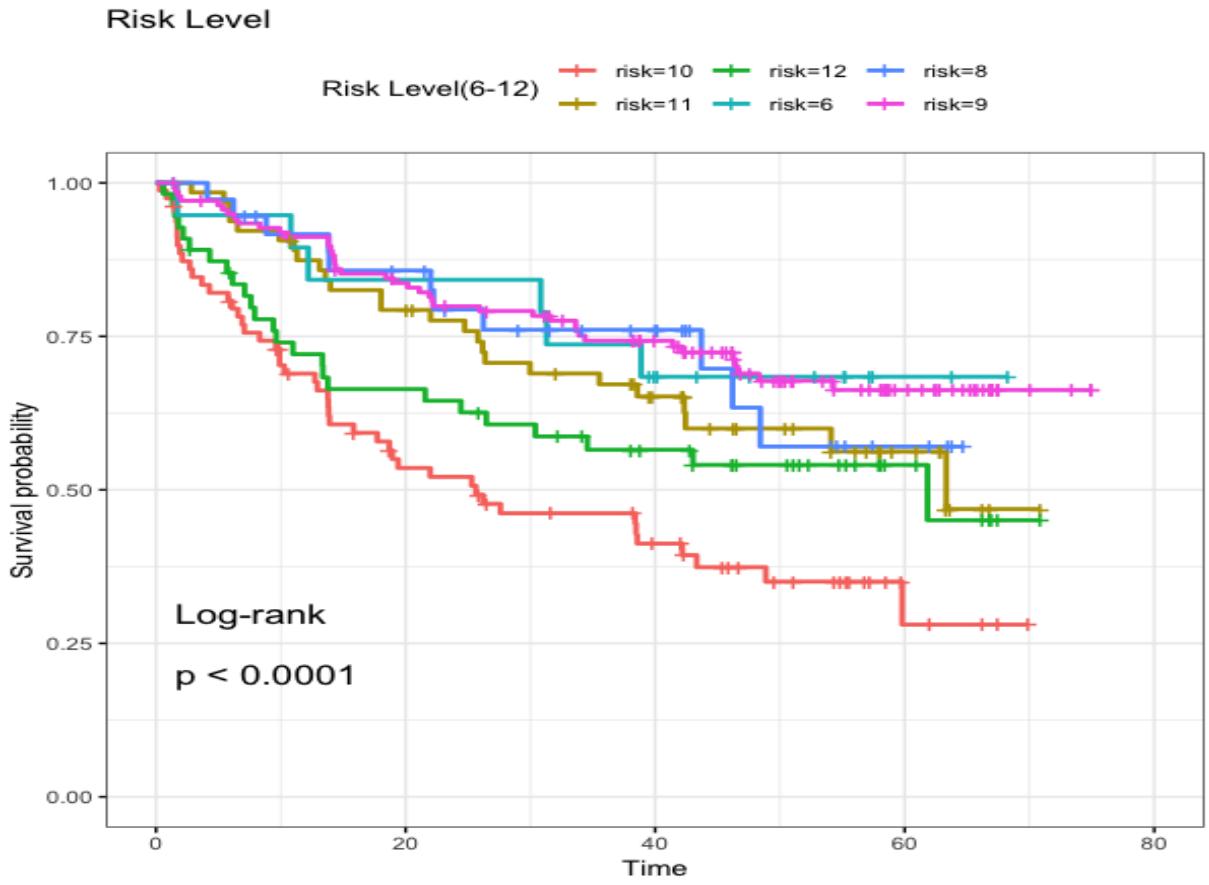


Figure 1 :

Figure 2 :

2.Cox PH Analysis Univariate

Cox-proportional hazards modeling can be used for single variables in order to estimate their effect on hazard rate. For instance, if we fit a cox model on laser, the model shows that the hazard proportion is 1.107 which indicates that the hazard rate for laser (Aragon) is closely similar to Xenon laser (difference of 0.107)

a. Laser

We conclude that with p-value of .3 (greater than .05) we say there is not a significant difference between the hazard rates for the patients that were treated with the two lasers.

b. Type

The p-value for our test is .9 which is greater than .05 so we conclude that there is no significant difference between the hazard rates of the patients with

juvenile diabetes than those with adult diabetes.

c. Eye
The p-value for this test is .09 which is slightly above .05 therefore we say there is not much significant difference between the hazard rate of patients that receive treatment on their right eye than those that receive treatment on their left eye.

d. Risk
With a p value less than .05 we conclude that there is a significant difference between the level of risk and the effect on patients with diabetes.

e. Treatment
From running a coxph test we see that the p value is 2e^-6 which is considerably less than 0.05 therefore we conclude that there is a significant difference between the patients that were in the treatment group than those that were in control.

**we fit the cox proportional hazard for each variable and the results are summarized on table1.

Variable	Exp(coef)	lower .95 : upper .95	p-value	likelihood	PH Assumption
Laser	1.175 (Aargon)	[0.8574 :1.61]	0.3	1	satisfied
Type	1.022 (Adult)	[0.744 :1.404]	0.9	0.02	satisfied
Eye	1.316 (Left)	[0.9524 :1.818]	0.09	2.89	satisfied
trt	0.46	[0.3304 :0.6403]	2e-06	22.35	satisfied
age	1.003	[0.9922 : 1.014]	0.6	0.27	satisfied
Risk8 Risk9 Risk10 Risk11 Risk12	1.0851 0.9614 2.8826 1.3758 1.8351	[0.4012-2.935] [0.4081-2.265] [1.2313- 6.748] [0.5654-3.361] [0.7526-4.474]	2e-06	28.95	satisfied

Table1: univariate cox proportional hazard

Conclusion: the covariate trt has the highest likelihood of 22.35 and p-value of 2e-06 which makes it significant and must include in our model.
The covariate risk also the highest likelihood of 22.35 and p-value of 2e-06 which makes it significant and must include in our model.

3.Covariates selection

By performing the backward selection and comparing the log likelihood of the covariates we see that age, risk, and treatments are the best covariates for our survival analysis because they have the highest log-likehood values and the smallest p-values as shown below.

```
> fitd <-coxph(Surv(ret.time, ret.status)~ret.age + ret.risk + ret.trt)
> anova(fitd)
```

```
##Analysis of Deviance Table
## Cox model: response is Surv(ret.time, ret.status)
## Terms added sequentially (first to last)

##         l    loglik     Chisq Df Pr(>|Chi|)
## NULL     -867.99
## ret.age  -822.78  90.416 53  0.0010370 **
## ret.risk  -811.62  22.323  5  0.0004546 ***
## ret.trt   -794.59  34.047  1  5.379e-09 ***
---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. CoxPH Multivariate Analysis

Testing all the covariates together, trt and risk10 are the most significant as show below

```
> cox.all = coxph(data = ret, ret.surv~ ret$laser + ret$eye + ret$trt +ret$age +ret$type + ret$risk)
   summary(cox.all)
```

```
##                      Pr(>|z|)
## ret$laserargon    0.2377
## ret$eyeleft       0.0934 .
## ret$trt           7.3e-07 ***
## ret$age           0.5368
## ret$typeadult     0.8645
## ret$risk8         0.9207
## ret$risk9         0.7317
## ret$risk10        0.0190 *
## ret$risk11        0.6618
## ret$risk12        0.2829
```

From our previous univariates analysis and covariate selection , we chose to test the covariates:

age, trt and type. Trt and risk 10 remains the most significant as shown below

```
> cox.risk = coxph(data = ret, ret.surv ~ ret$trt + ret$risk + ret$age)
> summary(cox.risk)
##                  Pr(>|z|)
## ret$trt        1.03e-06 ***
## ret$risk8       0.9458
## ret$risk9       0.8430
## ret$risk10     0.0147 *
## ret$risk11      0.5350
## ret$risk12      0.2221
## ret$age         0.5274
```

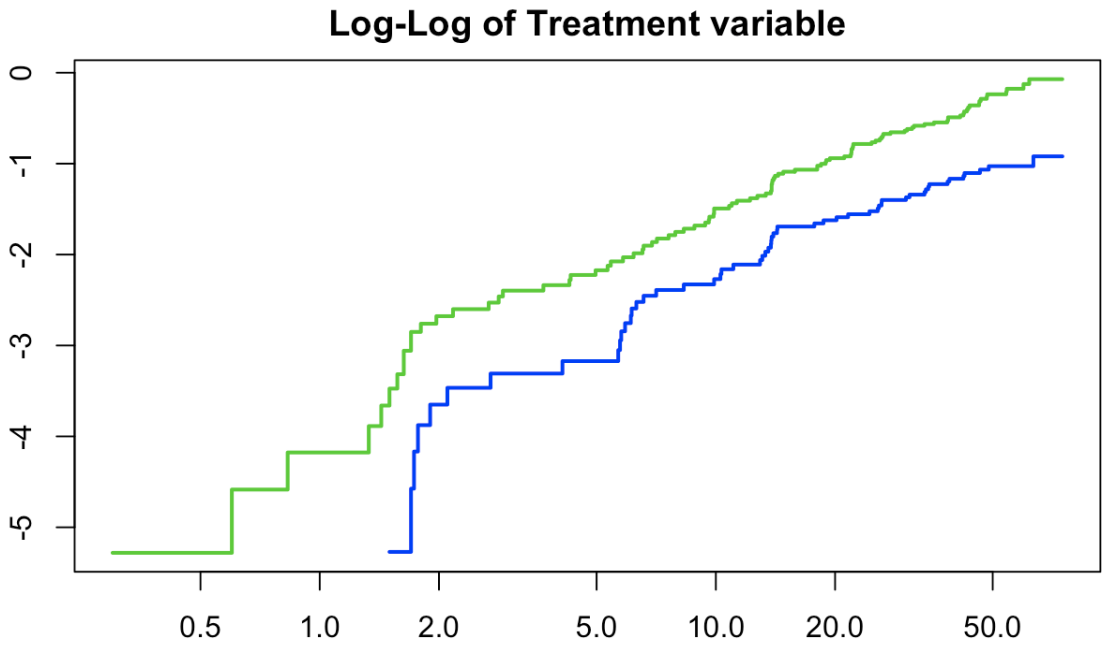
Model Checking

First we simply run a series of coxzph functions to test our Cox PH assumptions:

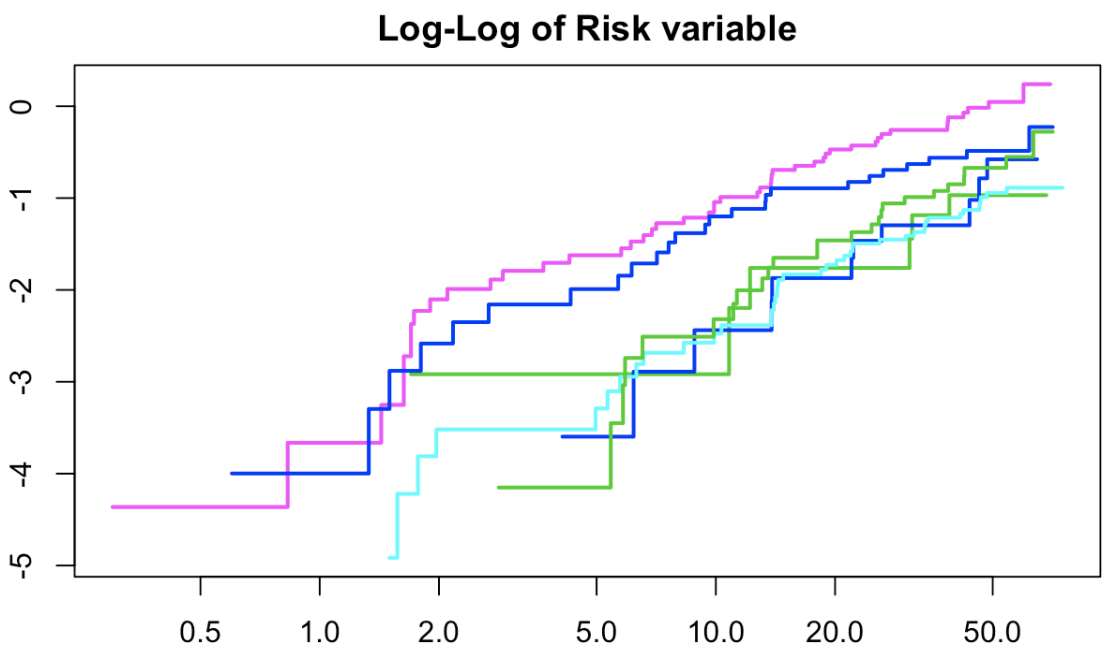
```
##                rho chisq     p
## laserargon -0.0764 0.898 0.343

##              rho chisq     p
## typeadult 0.0534 0.443 0.506

##            rho chisq     p
## eyeleft 0.0848  1.11 0.293

##             rho  chisq     p
## risk8    0.05107 0.4037 0.525
## risk9    0.00909 0.0128 0.910
## risk10 -0.03585 0.1975 0.657
## risk11   0.03880 0.2329 0.629
## risk12 -0.07470 0.8578 0.354
## GLOBAL       NA 8.1475 0.148

##         rho chisq     p
## trt -0.0564 0.495 0.482
```

None of our variables result in a p-value below .05 so we believe our assumptions are met.

Next we use a complementary log-log plot to visualize whether the Cox PH assumption is appropriate for modeling the effect of our factorized covariates.



We see from the Log-log plot for treatment the two curves are parallel so we can say that the assumptions hold, which agrees with our conclusion after performing the coxzph test.



This log-log plot of our risk variable seems to contradict our previous coxzph analysis.. Although it could be because we have a total of six levels, we still decide it would be best to look at stratifying the risk variable.

Stratification

After stratifying on the risk variable:

cox.strata.risk = coxph(data = ret, ret.surv ~ ret$trt + strata(ret$risk))

we find not too much of a difference in our tests. But we choose to include it in our model.

Paired observations are not independent. cluster()/frailty()

We have paired observations. For every subject there are two observations: one treated eye and one control eye. Because there is almost certainly correlation amongst each subject's eyes(each person has a different propensity of having their sight fail in either eye) we must 'cluster' the subject's two observations together to reduce the effect of a subject's eye correlation has on our analysis**.**

When we look at the results after using the cluster function on id against all covariates, we see that again only treatment and risk prove to be significant covariates in our analysis.
When we use the cluster id against laser, we get a p value a little bigger than .3 showing that even with modifying the data so as to account for the fact we have multiple observation for the same subject the type of laser has no significant effect on patients alone.
When we use cluster function against treatment, we again see that this covariate has significant effect on survival rate even with our adjustment to the data.

Our final model is:
coxph(data = ret, ret.surv ~ trt + strata(risk) + cluster(id))

If we run summary() on this model we can look at the confidence intervals of the hazard ratios

	exp(coef)	exp(-coef)	lower .95	upper .95
ret$trt	0.4478	2.233	0.3206	0.6254

For **trt** the 95% CI is (.3206, 6254)

Conclusions

The main takeaway from our analysis is that there is a distinct difference amongst the treated vs nontreated groups. Patients' treated eyes survived 44% longer until failure compared to their non-treated counterparts.

There is **no** difference in survival rates between left and right eyes.

Reference:

W. J. Huster, R. Brookmeyer and S. G. Self (1989). Modelling paired survival data with covariates, Biometrics 45:145-156.
A. L. Blair, D. R. Hadden, J. A. Weaver, D. B. Archer, P. B. Johnston and C. J. Maguire (1976). The 5-year prognosis for vision in diabetes, American Journal of Ophthalmology, 81:383-396.

Appendix

```
library(ggplot2)
library(tidyverse)

ret = retinopathy
```

#Exploratory Analysis/Prep Data

```{r}
names(ret)
head(ret)

#Factorize variables
ret = mutate(ret, risk = factor(risk))

#Create survival object
ret.surv = Surv(ret$futime, ret$status)

#Quantiles of survival times
###quantile(ret.surv, c(.25, .5, .75))

```

#Kaplan-Meier Analysis

```{r}
###Overall KM estimation###


#Fit to survival model
ret.survfit = survfit(ret.surv ~ 1, data = ret)
#Plot km estimate
dev.off()
ggsurvplot(data = ret, ret.survfit,
        legend = c(.7, .2),
        legend.title = 'Covariates',
        ggtheme = theme_bw(),
        title = 'All Covariates')


###KM of categorical variables###

ret.surv = Surv(ret$futime, ret$status)
#vs Laser type
km.laser = survfit(ret.surv ~ ret$laser)
ggsurvplot(data= ret, km.laser, conf.int = T, pval = T, pval.method = T,
        legend = c(.7, .2),
        legend.title = 'Laser Type',
        legend.labs = c('Argon', 'Xenon'),
        ggtheme = theme_bw(),
        title = ' Laser Type')
#vs type(age group)
km.type = survfit(ret.surv ~ ret$type)
ggsurvplot(data= ret, km.type, conf.int = T, pval = T, pval.method = T,
        legend = c(.7, .2),
        legend.title = 'Age Group',
        legend.labs = c('Adult', 'Juvenile'),
        ggtheme = theme_bw(),
        title = "Age Group")
#vs eye
km.eye = survfit(ret.surv ~ ret$eye)
ggsurvplot(data= ret, km.eye, conf.int = T, pval = T, pval.method = T,
        legend = c(.7, .2),
        legend.title = 'Eye',
        legend.labs = c('Left', 'Right'),
        ggtheme = theme_bw(),
        title = 'Eye(Left/Right)')
```

```
#vs risk
km.risk = survfit(ret.surv ~ ret$risk)
ggsurvplot(data= ret, km.risk, conf.int = F, pval = T, pval.method = T,
          legend.title = 'Risk Level(6-12)',
          ggtheme = theme_bw(),
          title = 'Risk Level')

#vs trt(treated/control)
km.trt = survfit(ret.surv ~ ret$trt)
ggsurvplot(data= ret, km.trt, conf.int = T, pval = T, pval.method = T,
          legend = c(.7, .2),
          legend.title = 'Treated or Controlled',
          legend.labs = c('Treated', 'Control'),
          ggtheme = theme_bw(),
          title = 'Treated vs Not treated')

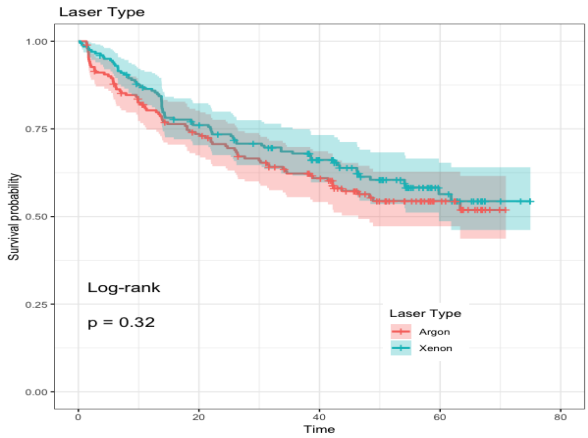
?retinopathy
retinopathy$age

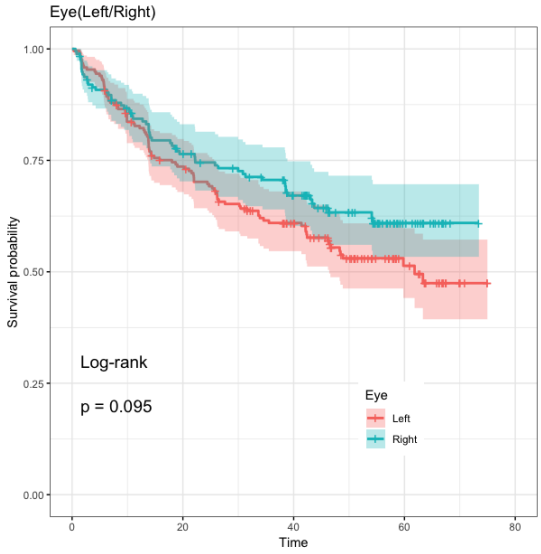
#CoxPH Analysis Univariate
```{r}
#Laser:
cox.laser = coxph(ret.surv ~ laser, data = ret)
summary(cox.laser)
# LRT, Wald, logrank: p  = 0.3
# exp(coef) = 1.175 (argon)
#INSIGNIFICIANT

#Type(adult/juve):
cox.type = coxph(ret.surv ~ type, data = ret)
summary(cox.type)
# LRT, Wald, logrank: p  = 0.9
# exp(coef) = 1.022 (adult)
#INSIGNIFICIANT

#Eye:
cox.eye = coxph(ret.surv ~ eye, data = ret)
summary(cox.eye)
# LRT, Wald, logrank: p  = 0.9
# exp(coef) = 1.316 (left)
#INSIGNIFICIANT

#Risk
cox.risk = coxph(ret.surv ~ risk, data = ret)
summary(cox.risk)
# LRT, Wald, logrank: p  = 0
#SIGNIFICIANT

#Treatment(treated/control)
cox.trt = coxph(ret.surv ~ trt, data = ret)
summary(cox.trt)
# LRT, Wald, logrank: p  = 0
# exp(coef) = 1.022
#SIGNIFICIANT

#Laser:
cox.age = coxph(ret.surv ~ age, data = ret)
summary(cox.age)
###Test coxph Assumptions with coxzph
cox.zph(cox.laser)
cox.zph(cox.type)
cox.zph(cox.eye)
cox.zph(cox.risk)
cox.zph(cox.trt)
cox.zph(cox.age)
#All pvalues > .05
#Assumptions HOLD

```
#Log-Log Plots
```{r}
```
```

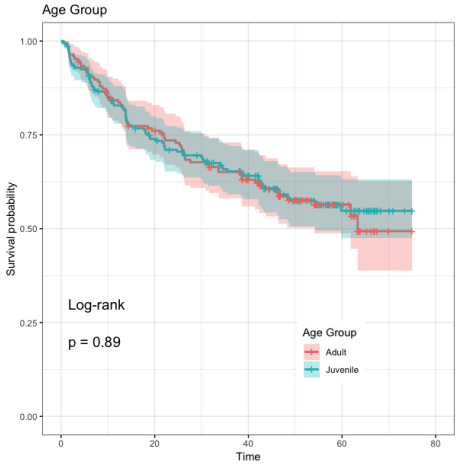
```
#Log-log of treatment variable
plot(survfit(ret.surv~ret$type), fun='cloglog', col=3:6,lwd=2,
     main = 'Log-Log of Treatment variable')
```

#Analysis of Clustered Observations

```{r}
#We cluster the id variable to create ambiguity of our paired observations
cluster.all = coxph(data = ret, ret.surv~ ret$laser + ret$eye + ret$trt + ret$type + ret$risk + cluster(id))
summary(cluster.all)

#Cluter id wagainst treatment
cluster.trt = coxph(data = ret, ret.surv ~ trt + cluster(id))
summary(cluster.trt)

#Cluster id against laser
cluster.laser = coxph(data = ret, ret.surv ~ laser + cluster(id))
summary(cluster.laser)


#Cluster id against trt and laser
cluster.risk = coxph(data = ret, ret.surv ~ age + laser + cluster(id))
summary(cluster.risk)
cox.trt_risk = coxph(data = ret, ret.surv ~ ret$trt + ret$risk)

# cluster id against trt and type
cluster.risk = coxph(data = ret, ret.surv ~ age + trt + type+ cluster(id))
summary(cluster.risk)
cox.trt_risk = coxph(data = ret, ret.surv ~ ret$trt + ret$risk)
```




Eye(Left/Right)
1.00
0.75
0.50
0.25
0.00
Survival probability
Log-rank
p = 0.095
Eye
Left
Right
0
20
40
60
80
Time


Age Group
1.00
0.75
0.50
0.25
0.00
Survival probability
Log-rank
p = 0.89
Age Group
Adult
Juvenile
0
20
40
60
80
Time