PSTAT 175 Final Project

Xiaoxi Guo, Zhuoqi Niu, Kelly Wang, David Zhang

12/5/2019

Load Dataset

```
NKI <- read.csv("NKI_cleaned.csv")
NKI <- NKI[,c('Patient','age','eventdeath','survival','timerecurrence','chemo','hormonal','amputation',
head(NKI)
```

```
##   Patient age eventdeath  survival timerecurrence chemo hormonal
## 1    s122  43          0 14.817248      14.817248     0        0
## 2    s123  48          0 14.261465      14.261465     0        0
## 3    s124  38          0  6.644764       6.644764     0        0
## 4    s125  50          0  7.748118       7.748118     0        1
## 5    s126  38          0  6.436687       6.318960     0        0
## 6    s127  42          0  5.037645       2.743326     1        0
##   amputation histtype diam posnodes grade angioinv
## 1          1        1   25        0     2        3
## 2          0        1   20        0     3        3
## 3          0        1   15        0     2        1
## 4          0        1   15        1     2        3
## 5          1        1   15        0     2        2
## 6          1        1   10        1     1        1
```

Load Packages

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

```
## Warning: package 'survminer' was built under R version 3.5.2
## Loading required package: ggplot2
## Loading required package: ggpubr
## Warning: package 'ggpubr' was built under R version 3.5.2
## Loading required package: magrittr
```

Data Exploration

```
summary(NKI[,c('age','survival','timerecurrence')])
```

```
##       age           survival       timerecurrence
##  Min.   :26.00   Min.   : 0.7118   Min.   : 0.271
##  1st Qu.:40.75   1st Qu.: 5.4997   1st Qu.: 4.389
##  Median :45.00   Median : 7.3593   Median : 6.950
```

```
##  Mean   :44.05   Mean   : 8.0806   Mean   : 7.250
##  3rd Qu.:49.00   3rd Qu.:10.5127   3rd Qu.: 9.986
##  Max.   :53.00   Max.   :18.3409   Max.   :18.341
```

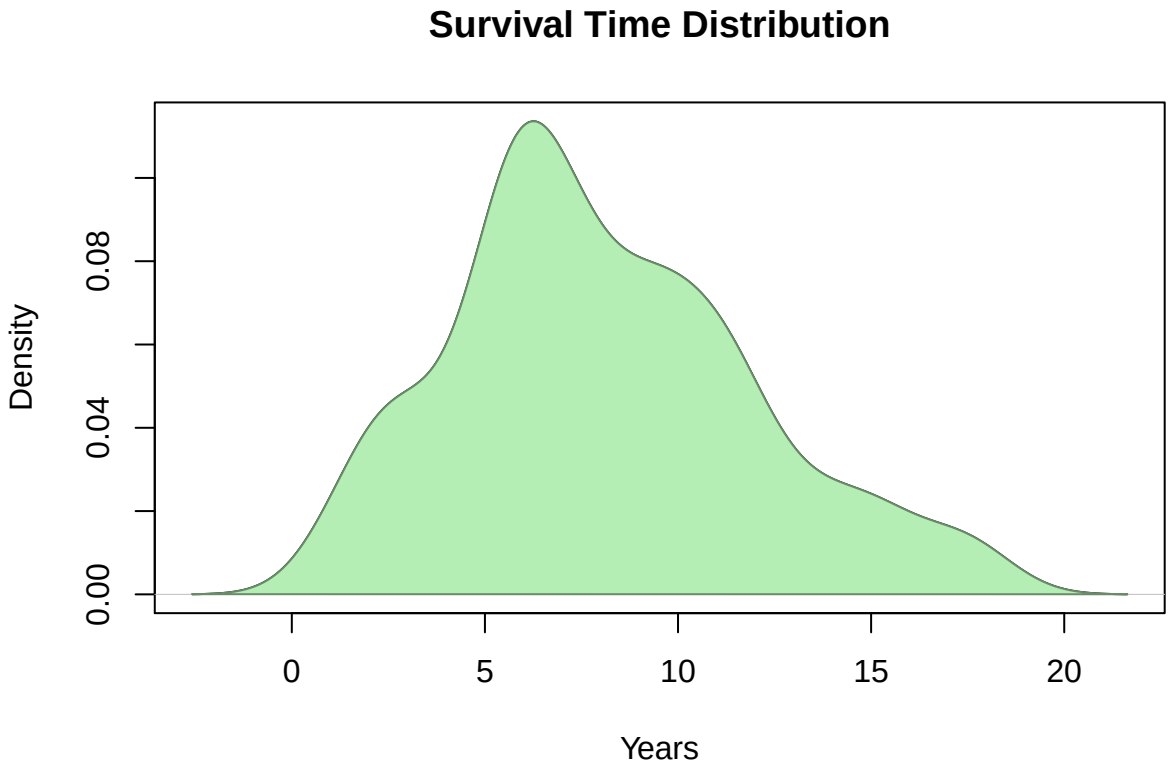
```
lapply(NKI[,c('chemo','hormonal','amputation','grade')], function(x) {
  return(table(x))
})
```

```
## $chemo
## x
##   0   1
## 165 107
##
## $hormonal
## x
##   0   1
## 236  36
##
## $amputation
## x
##   0   1
## 152 120
##
## $grade
## x
##   1   2   3
##  71  95 106
```

```
quantile(NKI$survival)
```

```
##        0%       25%       50%       75%      100%
##  0.711841  5.499738  7.359343 10.512662 18.340862
```

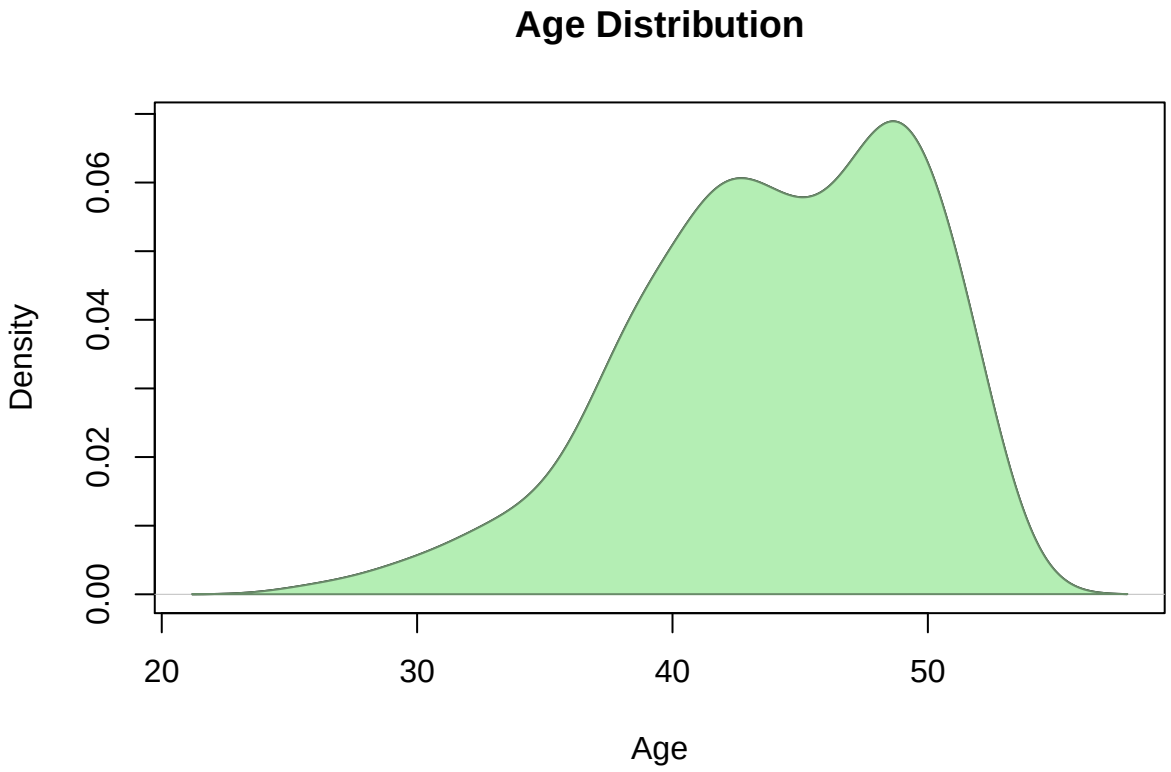
```
plot(density(NKI$survival), main='Survival Time Distribution', xlab='Years')
polygon(density(NKI$survival), col='darkseagreen2', border='darkseagreen4')
```




```
quantile(NKI$age)
```

```
##    0%   25%   50%   75%  100%
## 26.00 40.75 45.00 49.00 53.00
```

```
plot(density(NKI$age), main='Age Distribution', xlab='Age')
polygon(density(NKI$age), col='darkseagreen2', border='darkseagreen4')
```



Group by approximately 25% and 75% quantile

```
NKI$diamgroup[NKI$diam<=15] = 1
NKI$diamgroup[NKI$diam>15 & NKI$diam<30] = 2
NKI$diamgroup[NKI$diam>=30] = 3

NKI$agegroup[NKI$age<=40.75] = 1
NKI$agegroup[NKI$age>40.75 & NKI$age<49.00] = 2
NKI$agegroup[NKI$age>=49.00] = 3
```

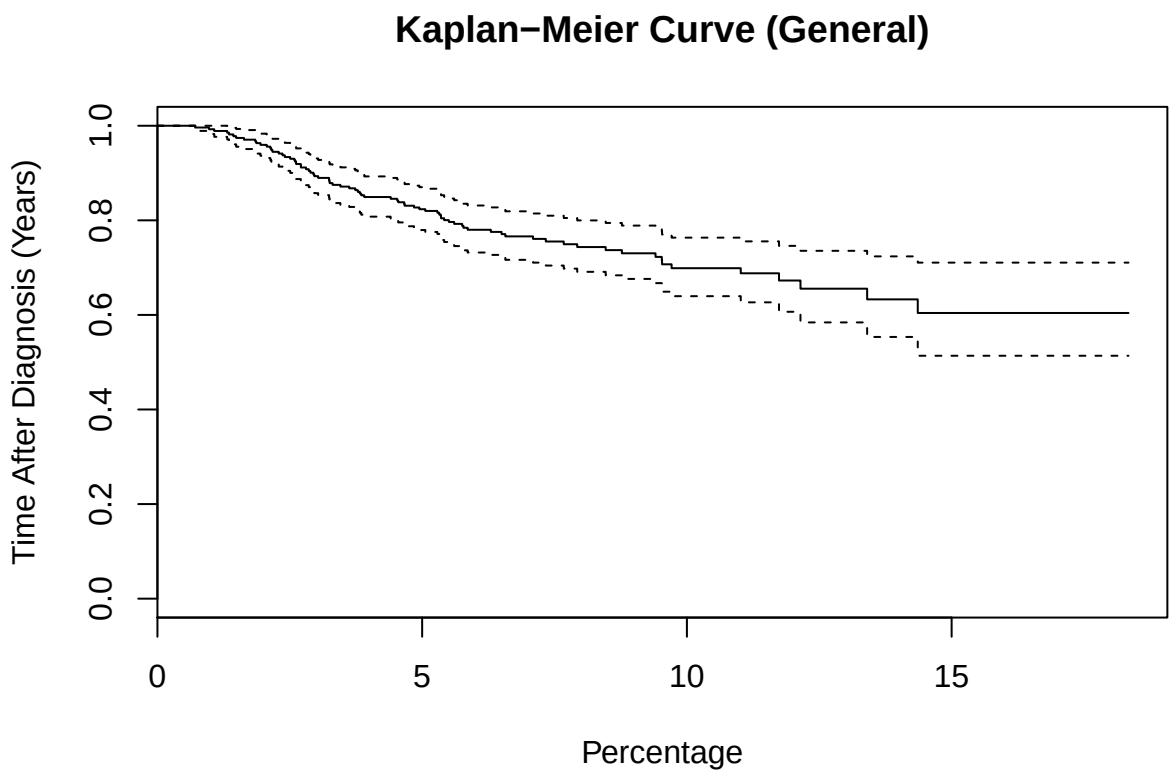
```
head(NKI) #Check first rows of dataset
```

```
##   Patient age eventdeath  survival timerecurrence chemo hormonal
## 1    s122  43          0 14.817248      14.817248     0        0
## 2    s123  48          0 14.261465      14.261465     0        0
## 3    s124  38          0  6.644764       6.644764     0        0
## 4    s125  50          0  7.748118       7.748118     0        1
## 5    s126  38          0  6.436687       6.318960     0        0
## 6    s127  42          0  5.037645       2.743326     1        0
##   amputation histtype diam posnodes grade angioinv diamgroup agegroup
## 1          1        1   25        0     2        3         2        2
## 2          0        1   20        0     3        3         2        2
## 3          0        1   15        0     2        1         1        1
## 4          0        1   15        1     2        3         1        3
## 5          1        1   15        0     2        2         1        1
## 6          1        1   10        1     1        1         1        2
```

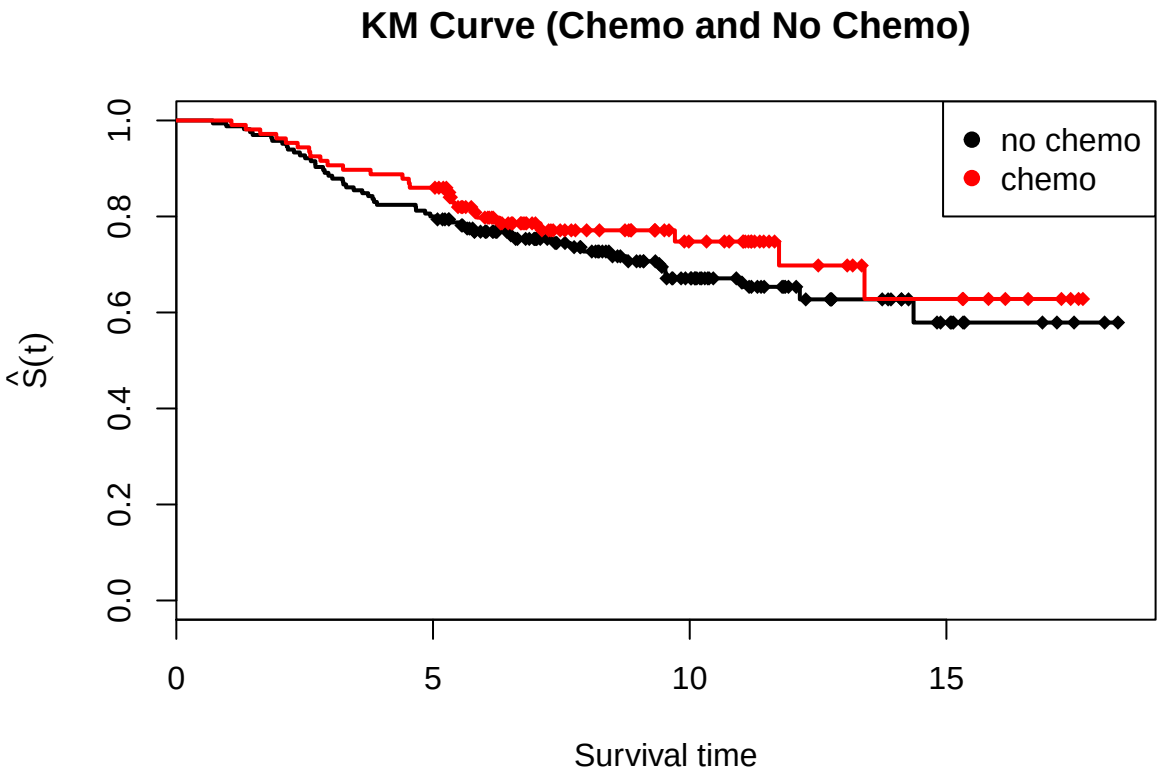
Covariates that we are interested are *chemo*, *hormonal*, *amputation*, *diamgroup*, *agegroup*, *grade*.

KM Plots

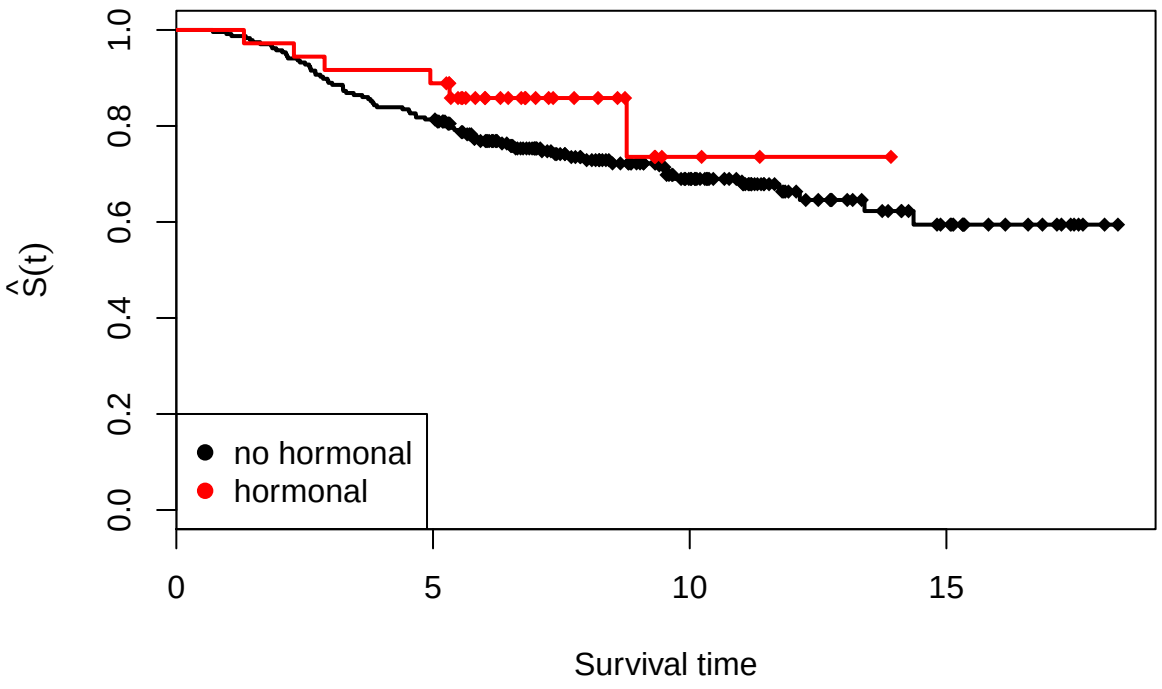
```
NKI.surv <- Surv(NKI$survival,NKI$eventdeath)
NKI.fit <- survfit(NKI.surv ~ 1)
plot(NKI.fit, main = "Kaplan-Meier Curve (General)", xlab = "Percentage", ylab = "Time After Diagnosis
```



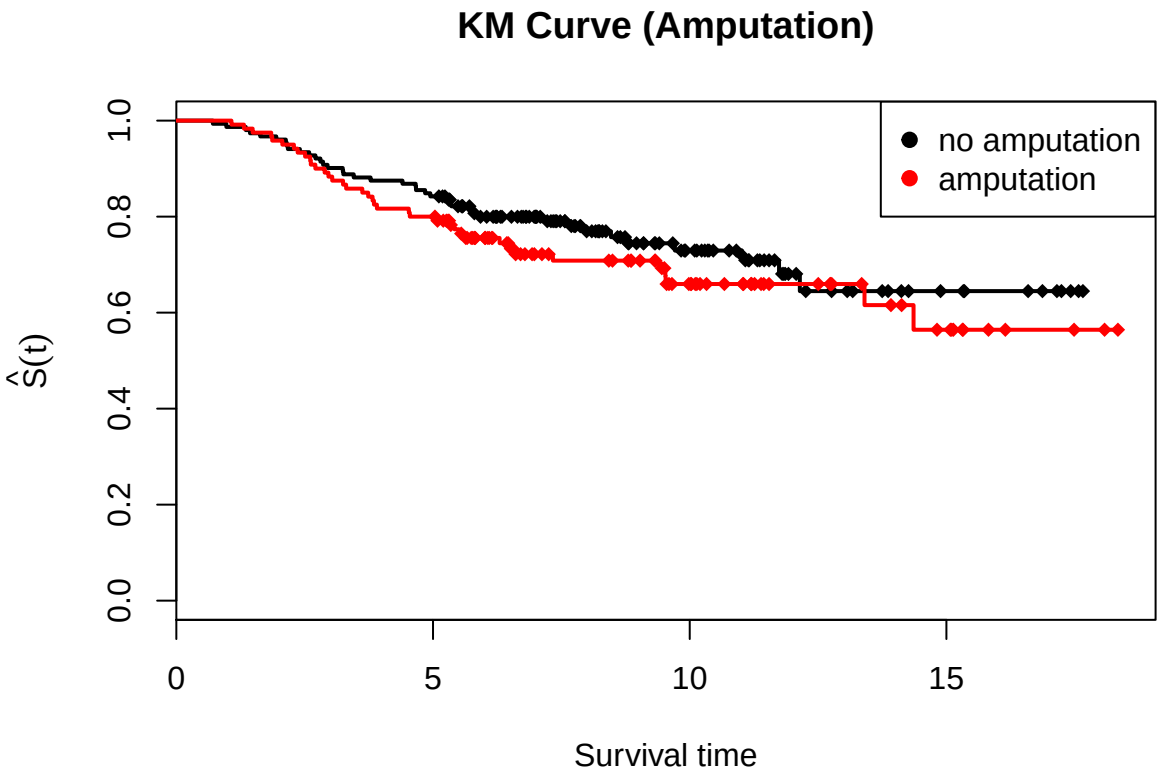

```
km_chemo = survfit(NKI.surv~chemo,data=NKI)
par(mar=c(5,5,4,2))
plot(km_chemo, xlab="Survival time",ylab = expression(hat(S)(t)),main = "KM Curve (Chemo and No Chemo)"
legend("topright",legend=c("no chemo","chemo"),col=1:2,pch=rep(19,2))
```



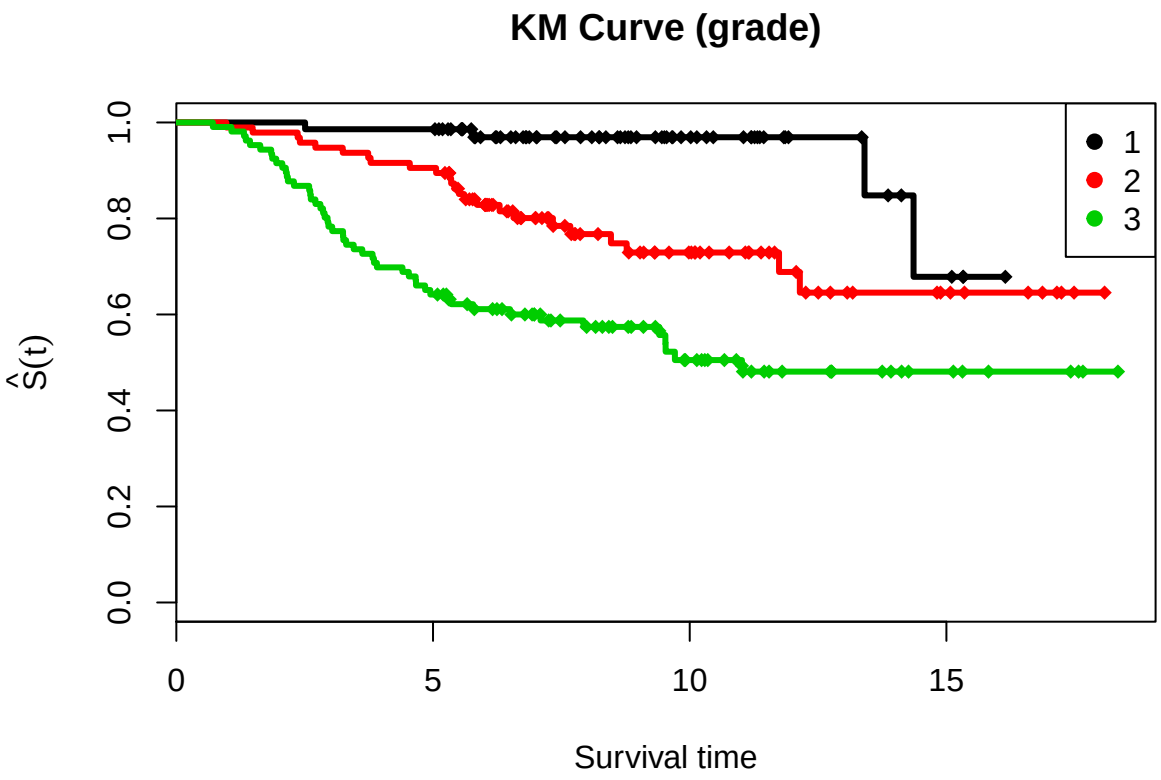

```
km_hormonal = survfit(NKI.surv~hormonal,data=NKI)
par(mar=c(5,5,4,2))
plot(km_hormonal, xlab="Survival time",ylab = expression(hat(S)(t)),main = "KM Curve (hormonal)",lwd=2,
legend("bottomleft",legend=c("no hormonal","hormonal"),col=1:2,pch=rep(19,2))
```



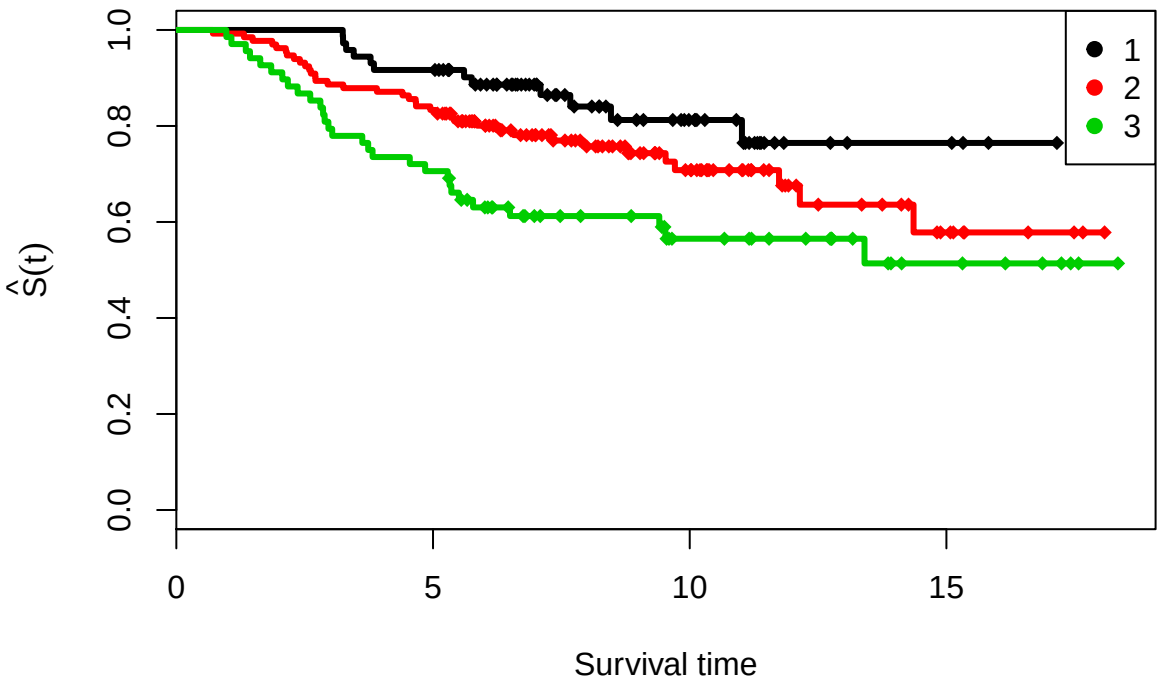
```
km_amputation = survfit(NKI.surv~amputation,data=NKI)
par(mar=c(5,5,4,2))
plot(km_amputation, xlab="Survival time",ylab = expression(hat(S)(t)),main = "KM Curve (Amputation)",lwd
legend("topright",legend=c("no amputation","amputation"),col=1:2,pch=rep(19,2))
```



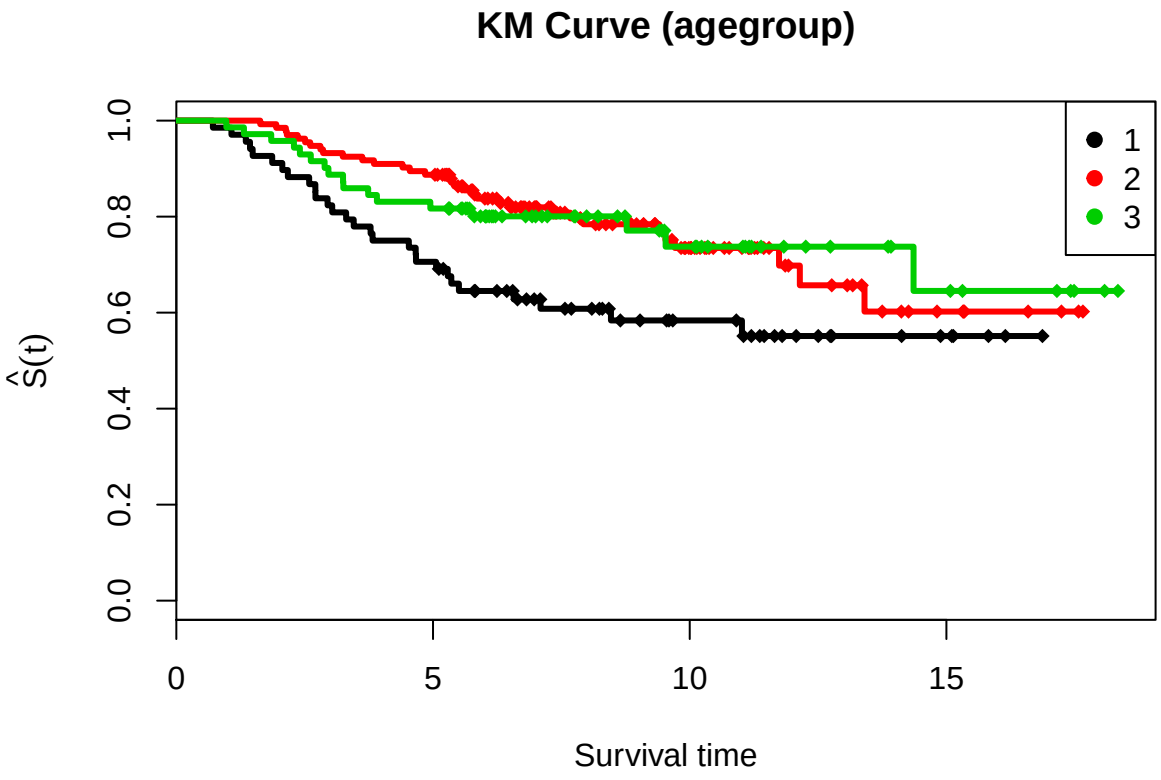

```
km_grade = survfit(NKI.surv~grade,data=NKI)
par(mar=c(5,5,4,2))
plot(km_grade, xlab="Survival time",ylab = expression(hat(S)(t)),main = "KM Curve (grade)",lwd=3, col=1
legend("topright",legend=c("1","2","3"),col=1:3,pch=rep(19,2))
```




```
km_diamgroup = survfit(NKI.surv~diamgroup,data=NKI)
par(mar=c(5,5,4,2))
plot(km_diamgroup, xlab="Survival time",ylab = expression(hat(S)(t)),main = "KM Curve (diamgroup)",lwd=
legend("topright",legend=c("1","2","3"),col=1:3,pch=rep(19,2))
```



```r
km_agegroup = survfit(NKI.surv~agegroup,data=NKI)
par(mar=c(5,5,4,2))
plot(km_agegroup, xlab="Survival time",ylab = expression(hat(S)(t)),main = "KM Curve (agegroup)",lwd=3,
legend("topright",legend=c("1","2","3"),col=1:3,pch=rep(19,2))
```



LogRank Test

```r
survdiff(NKI.surv~NKI$chemo)
```

```
## Call:
## survdiff(formula = NKI.surv ~ NKI$chemo)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## NKI$chemo=0 165       51     46.8     0.378     0.964
## NKI$chemo=1 107       26     30.2     0.585     0.964
##
##  Chisq= 1  on 1 degrees of freedom, p= 0.3
```

```r
survdiff(NKI.surv~NKI$hormonal)
```

```
## Call:
## survdiff(formula = NKI.surv ~ NKI$hormonal)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## NKI$hormonal=0 236       71    67.92     0.139       1.2
## NKI$hormonal=1  36        6     9.08     1.044       1.2
##
##  Chisq= 1.2  on 1 degrees of freedom, p= 0.3
```

```
survdiff(NKI.surv~NKI$amputation)
```

```
## Call:
## survdiff(formula = NKI.surv ~ NKI$amputation)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## NKI$amputation=0 152       39     43.5     0.460      1.06
## NKI$amputation=1 120       38     33.5     0.596      1.06
##
##  Chisq= 1.1  on 1 degrees of freedom, p= 0.3
```

```
survdiff(NKI.surv~NKI$grade)
```

```
## Call:
## survdiff(formula = NKI.surv ~ NKI$grade)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## NKI$grade=1   71        4     22.3    15.032     21.20
## NKI$grade=2   95       24     28.4     0.672      1.07
## NKI$grade=3 106       49     26.3    19.549     29.79
##
##  Chisq= 35.4  on 2 degrees of freedom, p= 2e-08
```

```
survdiff(NKI.surv~NKI$agegroup)
```

```
## Call:
## survdiff(formula = NKI.surv ~ NKI$agegroup)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## NKI$agegroup=1   68       28     17.7     5.963     7.754
## NKI$agegroup=2 133       32     39.1     1.294     2.638
## NKI$agegroup=3  71       17     20.2     0.497     0.674
##
##  Chisq= 7.8  on 2 degrees of freedom, p= 0.02
```

```
survdiff(NKI.surv~NKI$diamgroup)
```

```
## Call:
## survdiff(formula = NKI.surv ~ NKI$diamgroup)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## NKI$diamgroup=1   72       12     21.4    4.1479     5.774
## NKI$diamgroup=2 132       36     37.4    0.0531     0.103
## NKI$diamgroup=3  68       29     18.2    6.4657     8.519
##
##  Chisq= 10.8  on 2 degrees of freedom, p= 0.005
```

Model Selection

Model1: Backward selection

```
cox <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup+grade+agegroup, data = NKI)
step(cox, direction = "backward")
```

```
## Start:  AIC=768.56
## Surv(NKI$survival, NKI$eventdeath) ~ diamgroup + grade + agegroup
##
##             Df    AIC
## <none>         768.56
## - agegroup   1 770.02
## - diamgroup  1 770.76
## - grade      1 795.21
```

```
## Call:
## coxph(formula = Surv(NKI$survival, NKI$eventdeath) ~ diamgroup +
##     grade + agegroup, data = NKI)
##
##              coef exp(coef) se(coef)      z        p
## diamgroup  0.3347    1.3976   0.1650  2.029   0.0424
## grade      0.8938    2.4445   0.1842  4.852 1.22e-06
## agegroup  -0.2988    0.7417   0.1617 -1.848   0.0646
##
## Likelihood ratio test=44.81  on 3 df, p=1.013e-09
## n= 272, number of events= 77
```

Model2: Likelihood tests selection

```
cox1 <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup+grade+agegroup, data = NKI)
anova(cox1)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(NKI$survival, NKI$eventdeath)
## Terms added sequentially (first to last)
##
##            loglik   Chisq Df Pr(>|Chi|)
## NULL      -403.69
## diamgroup -398.42 10.5380  1   0.001169 **
## grade     -383.01 30.8137  1   2.84e-08 ***
## agegroup  -381.28  3.4623  1   0.062784 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All variables are significant (using 0.1 as critical value).

Compared model

```
fit0 <- coxph(Surv(survival, eventdeath)~grade, data = NKI)
fit1 <- coxph(Surv(survival, eventdeath)~grade+diamgroup, data = NKI)
lrt1 = 2*(fit1$loglik[2]-fit0$loglik[2])
pchisq(lrt1,df=1,lower.tail = FALSE)
```

```
## [1] 0.06125396
```

```
fit2 <- coxph(Surv(survival, eventdeath)~grade+diamgroup+agegroup, data = NKI)
lrt2 = 2*(fit2$loglik[2]-fit1$loglik[2])
pchisq(lrt2,df=1,lower.tail = FALSE)
```

```
## [1] 0.06278422
```

Do not mention the specific numbers from the compared model part; Give our final decision only since the numbers do not look beautiful.

Model Checking

Method1: Residual tests

```
NKI <- within(NKI, {
  grade <- factor(grade, labels = c("1", "2","3"))
  diamgroup <- factor(diamgroup, labels = c("1", "2","3"))
  agegroup <- factor(agegroup, labels = c("1", "2","3"))
})

cox1 <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup+grade+agegroup, data = NKI)

cox.zph(cox1)
```

```
##                 rho    chisq      p
## diamgroup2 -0.14656  1.67725 0.1953
## diamgroup3 -0.21738  3.57490 0.0587
## grade2     -0.13849  1.51140 0.2189
## grade3     -0.23006  4.00697 0.0453
## agegroup2   0.23904  4.36662 0.0366
## agegroup3   0.00654  0.00326 0.9545
## GLOBAL           NA 16.46216 0.0115
```

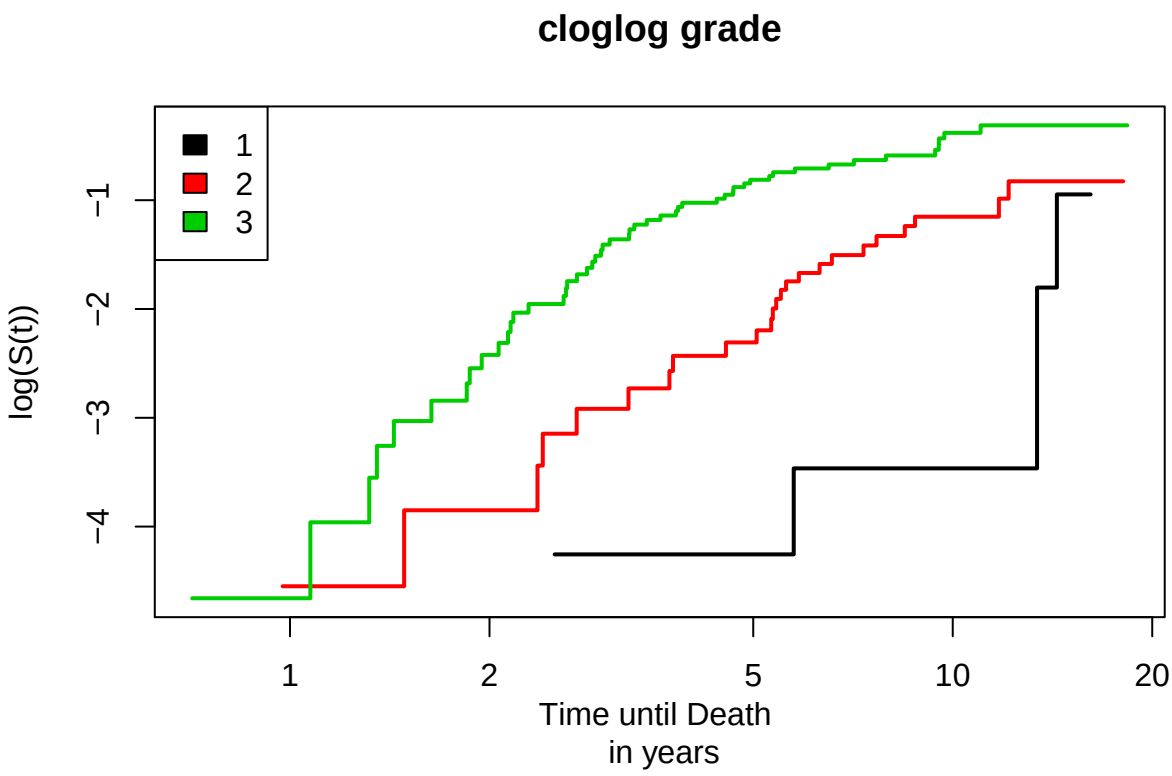
Since p-value for *grade* is less than 0.05, we need to stratify it.

```
cox2 <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup+strata(grade)+agegroup, data = NKI)
cox.zph(cox2)
```

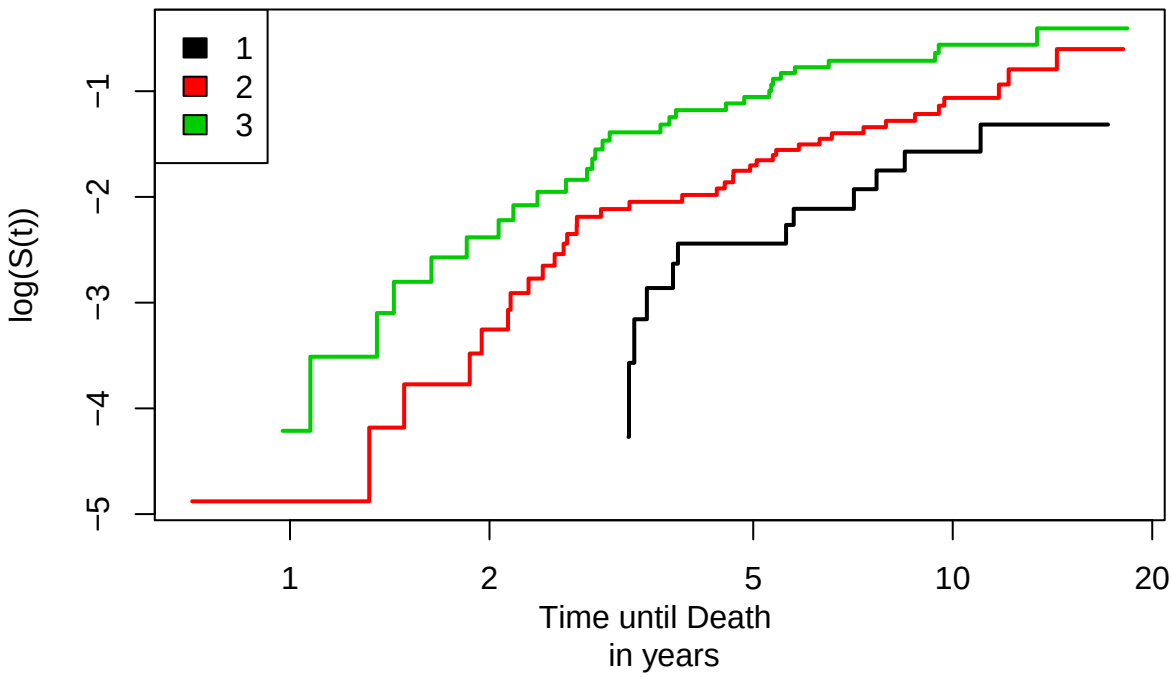
```
##                rho  chisq      p
## diamgroup2 -0.1443 1.6146 0.2038
## diamgroup3 -0.2136 3.3979 0.0653
## agegroup2   0.2620 5.2164 0.0224
## agegroup3   0.0292 0.0663 0.7969
## GLOBAL          NA 8.3750 0.0788
```

Method2:C-log-log Plot

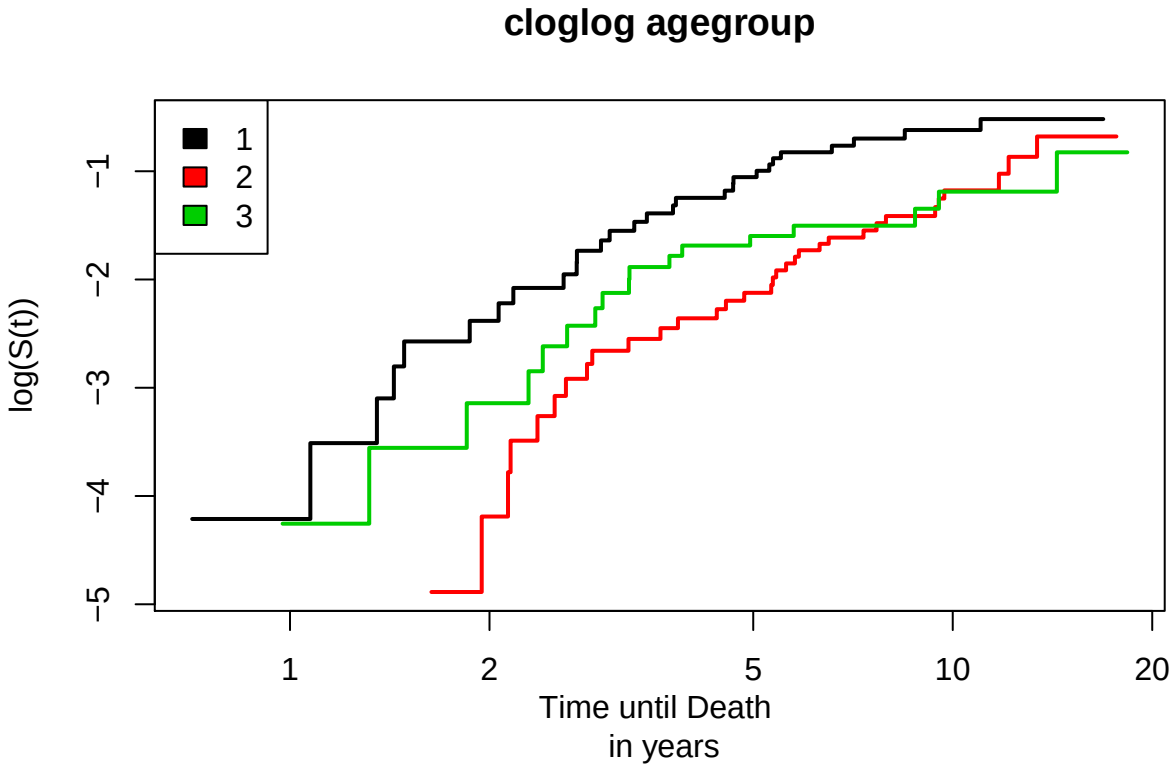
```
#grade
plot(survfit(NKI.surv ~ NKI$grade),lwd=2,col=1:3, fun="cloglog",main = "cloglog grade",xlab="Time until
legend('topleft',c("1","2","3"),fill = 1:3)
```




```
#diamgroup
plot(survfit(NKI.surv ~ NKI$diamgroup),lwd=2,col=1:3, fun="cloglog",main = "cloglog diamgroup",xlab="Ti
legend('topleft',c("1","2","3"),fill = 1:3)
```



```r
#agegroup
plot(survfit(NKI.surv ~ NKI$agegroup),lwd=2,col=1:3, fun="cloglog",main = "cloglog agegroup",xlab="Time
legend('topleft',c("1","2","3"),fill = 1:3)
```



Interaction Terms

```r
coxA <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup*strata(agegroup), data = NKI)
coxB <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup*strata(grade), data = NKI)
coxC <- coxph(Surv(NKI$survival, NKI$eventdeath)~strata(agegroup)*strata(grade), data = NKI)
```

```
## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 1,2,3 ; beta may be infinite.
```

```r
anova(coxA)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(NKI$survival, NKI$eventdeath)
## Terms added sequentially (first to last)
##
##                              loglik  Chisq Df Pr(>|Chi|)
## NULL                        -318.38
## diamgroup                   -312.27 12.208  2   0.002234 **
## diamgroup:strata(agegroup) -310.86  2.813  4   0.589594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(coxB)
```

```
## Analysis of Deviance Table
```

```
##  Cox model: response is Surv(NKI$survival, NKI$eventdeath)
## Terms added sequentially (first to last)
##
##                           loglik  Chisq Df Pr(>|Chi|)
## NULL                     -322.71
## diamgroup                -320.63 4.1711  2     0.1242
## diamgroup:strata(grade) -320.43 0.3944  4     0.9829
```

```
anova(coxC)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(NKI$survival, NKI$eventdeath)
## Terms added sequentially (first to last)
##
##                                 loglik Chisq Df Pr(>|Chi|)
## NULL                           -240.06
## strata(agegroup):strata(grade) -240.06     0  3          1
```

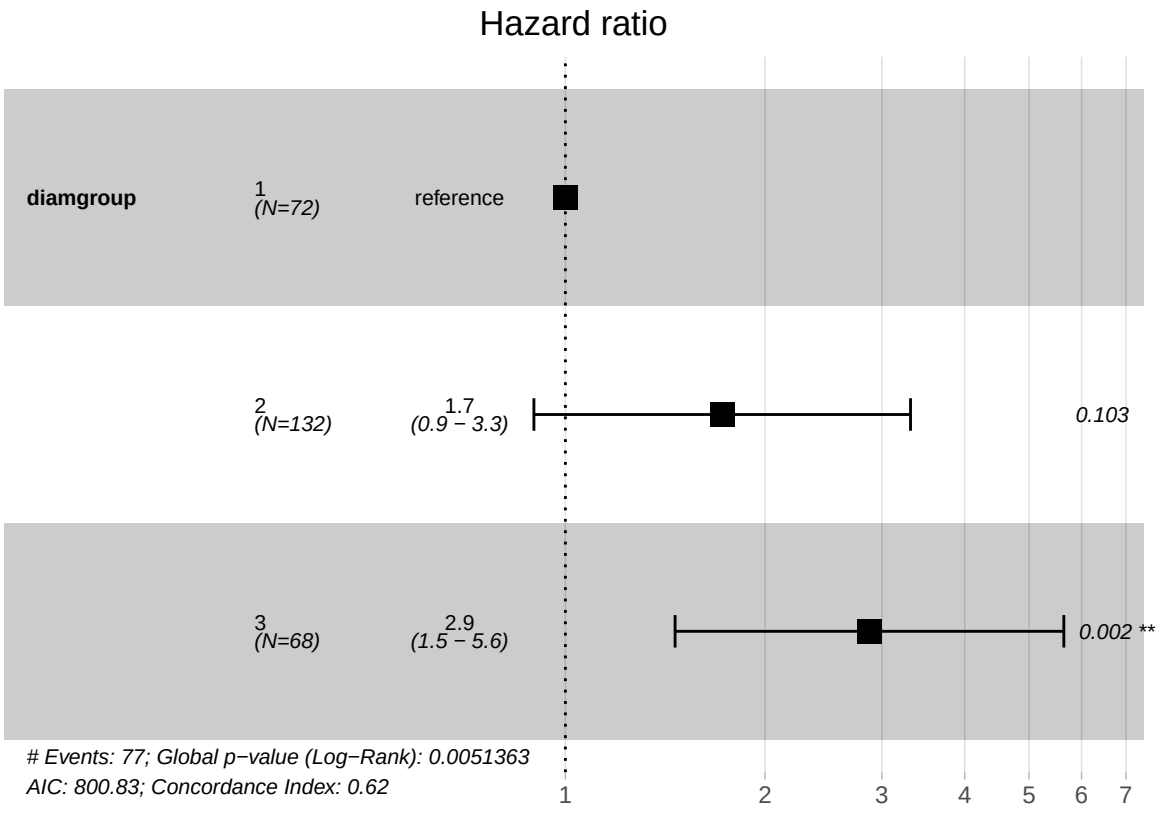
All the interaction terms have large p-values (greater than 0.05). Therefore, we will not include interaction terms in our model.

Hazard Ratios and C.I.

```
cox_ns <- coxph(Surv(NKI$survival, NKI$eventdeath)~diamgroup, data = NKI)
ggforest(cox_ns, data = NULL, main = "Hazard ratio",
         cpositions = c(0.02, 0.22, 0.4), fontsize = 0.7,
         refLabel = "reference", noDigits = 2)
```

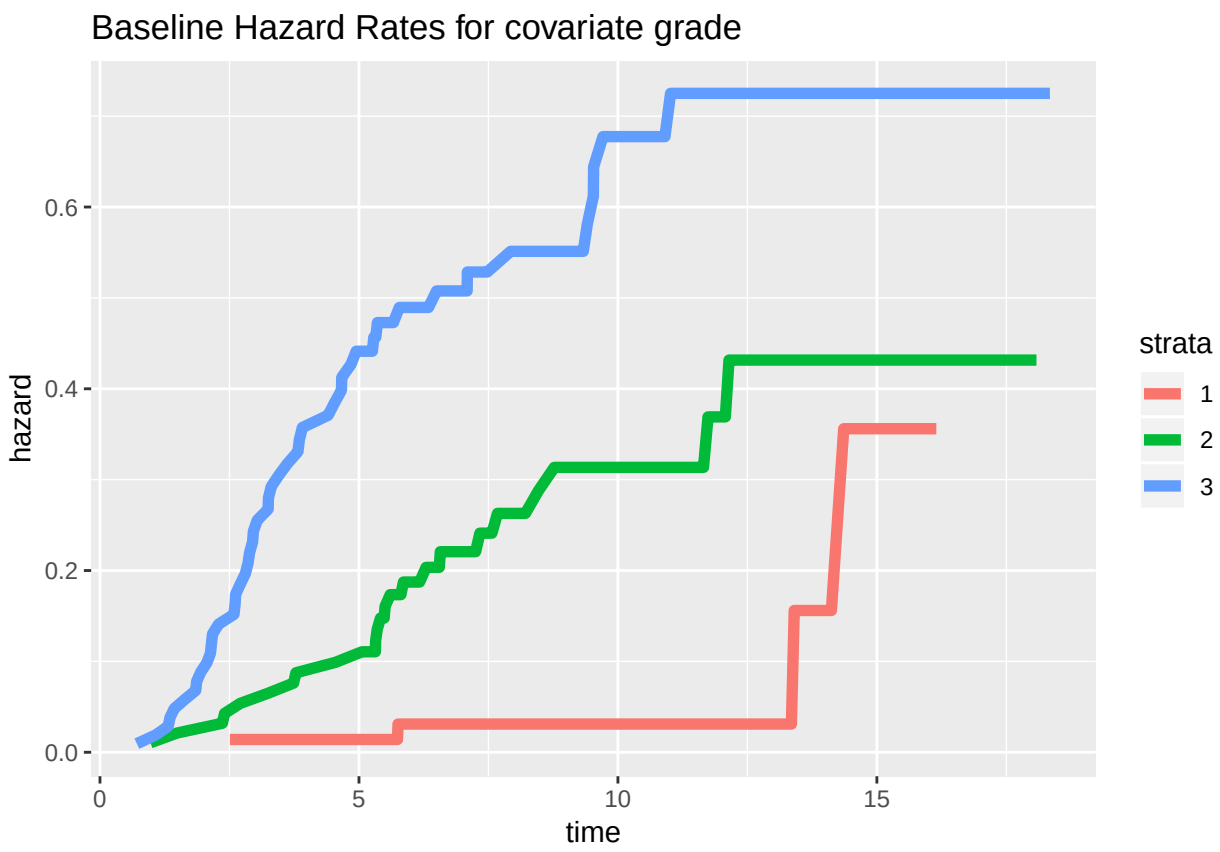
```
## Warning in .get_data(model, data = data): The `data` argument is not
## provided. Data will be extracted from model fit.
```

```
## Warning: Removed 1 rows containing missing values (geom_errorbar).
```

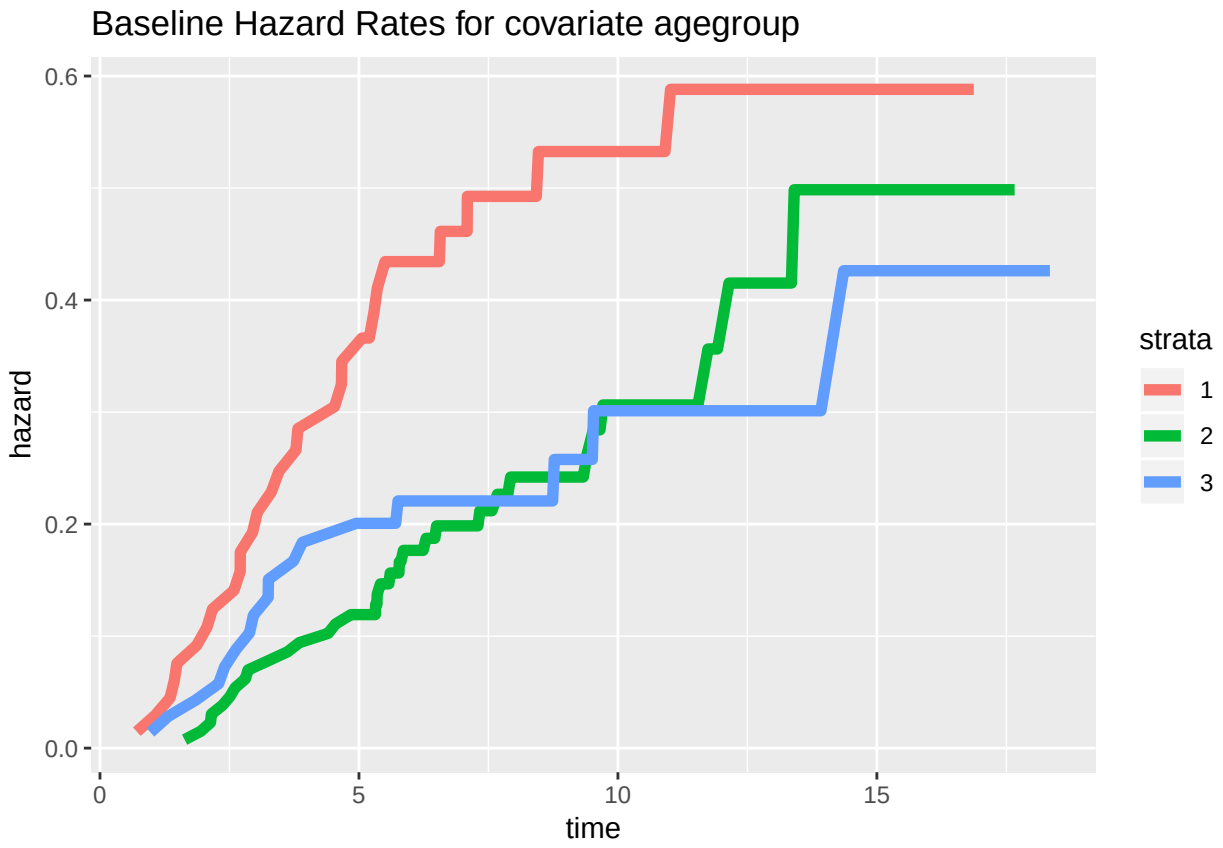


Baseline Hazard Rates

```
#Baseline Hazard Plot for grade
fit_grade <- coxph(Surv(survival, eventdeath)~strata(grade), data=NKI)
by_grade <- basehaz(fit_grade) %>%
  group_by(strata)
ggplot(by_grade, aes(x = time, y = hazard)) +
  geom_line(aes(color = strata), size=2) +
  ggtitle("Baseline Hazard Rates for covariate grade")
```




```
#Baseline Hazard Plot for agegroup
fit_age <- coxph(Surv(survival, eventdeath)~strata(agegroup), data=NKI)
by_age <- basehaz(fit_age) %>%
  group_by(strata)
ggplot(by_age, aes(x = time, y = hazard)) +
  geom_line(aes(color = strata), size=2) +
  ggtitle("Baseline Hazard Rates for covariate agegroup")
```


Baseline Hazard Rates for covariate agegroup
hazard
0.0
0.2
0.4
0.6
0
5
10
15
time
strata
1
2
3