Data Analysis

Juste Simanauskaite & Patricia Rivera

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3, fig.width=5, fig.align="center")
library(tidyverse)
library(broom)
library(plyr)
library(survival)
library(survminer)
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")
dim(aids)
```

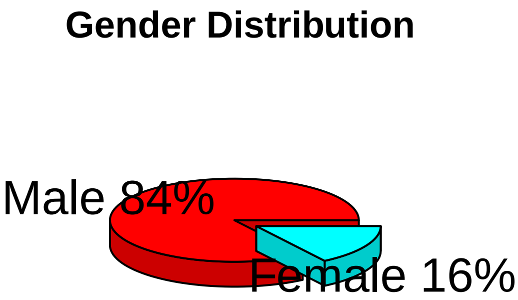
```
## [1] 851  16
```

```
summary(aids)
```

```
##        id              time           censor            time_d
##  Min.   :   1.0   Min.   :  1.0   Min.   :0.00000   Min.   :  1.0
##  1st Qu.: 287.5   1st Qu.:179.5   1st Qu.:0.00000   1st Qu.:199.5
##  Median : 581.0   Median :257.0   Median :0.00000   Median :266.0
##  Mean   : 579.5   Mean   :231.8   Mean   :0.08108   Mean   :243.4
##  3rd Qu.: 873.0   3rd Qu.:300.0   3rd Qu.:0.00000   3rd Qu.:306.0
##  Max.   :1156.0   Max.   :362.0   Max.   :1.00000   Max.   :362.0
##     censor_d            tx             txgrp           strat2
##  Min.   :0.0000   Min.   :0.0000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :2.000   Median :1.0000
##  Mean   :0.0235   Mean   :0.5041   Mean   :1.504   Mean   :0.6157
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :2.000   Max.   :1.0000
##       sex            raceth          ivdrug         hemophil
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :0.00000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.00000
##  Median :1.000   Median :1.000   Median :1.000   Median :0.00000
##  Mean   :1.157   Mean   :1.706   Mean   :1.317   Mean   :0.03408
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.00000
##  Max.   :2.000   Max.   :5.000   Max.   :3.000   Max.   :1.00000
##      karnof            cd4            priorzdv           age
##  Min.   : 70.00   Min.   :  0.00   Min.   :  3.00   Min.   :15.00
##  1st Qu.: 90.00   1st Qu.: 22.25   1st Qu.: 11.00   1st Qu.:33.00
##  Median : 90.00   Median : 75.00   Median : 21.00   Median :38.00
##  Mean   : 91.34   Mean   : 86.45   Mean   : 30.63   Mean   :38.81
##  3rd Qu.:100.00   3rd Qu.:135.75   3rd Qu.: 44.00   3rd Qu.:44.00
##  Max.   :100.00   Max.   :348.00   Max.   :288.00   Max.   :73.00
```

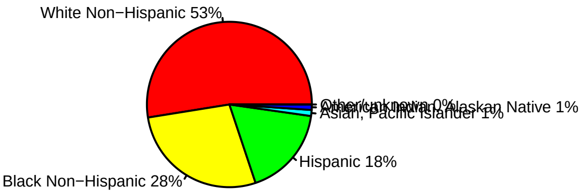
The data set contains a sample size equal to 851 participants and 16 different variables.

```
library(plotrix)
male<-sum(aids$sex==1)
female<-sum(aids$sex==2)
slices <- c(male, female)
lbls <- c("Male", "Female")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie3D(slices,labels=lbls,explode=0.1,
   main="Gender Distribution ", cex.lab=0.1)
```

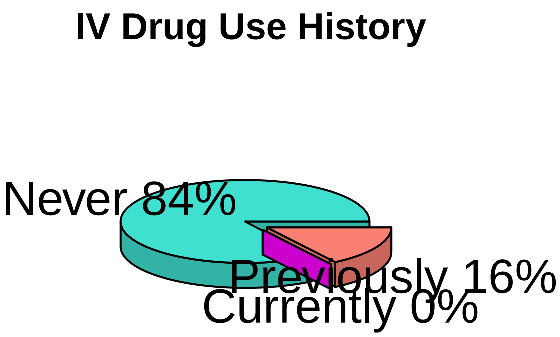


The Pie Chart represents the gender distribution in the sample, with 84% male and 16% female.

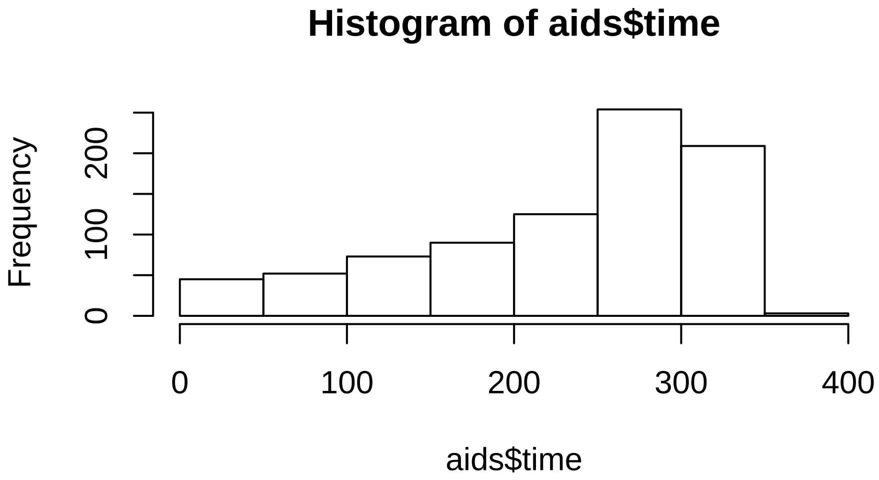
```
wnh<-sum(aids$raceth==1)
bnh<-sum(aids$raceth==2)
h<-sum(aids$raceth==3)
api<-sum(aids$raceth==4)
aian<-sum(aids$raceth==5)
oth<-sum(aids$raceth==6)
slices <- c(wnh,bnh,h,api,aian,oth)
lbls <- c("White Non-Hispanic", "Black Non-Hispanic", "Hispanic","Asian, Pacific Islander", "American I
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie(slices,lbls,col = rainbow(length(lbls)), cex=0.5 )
```

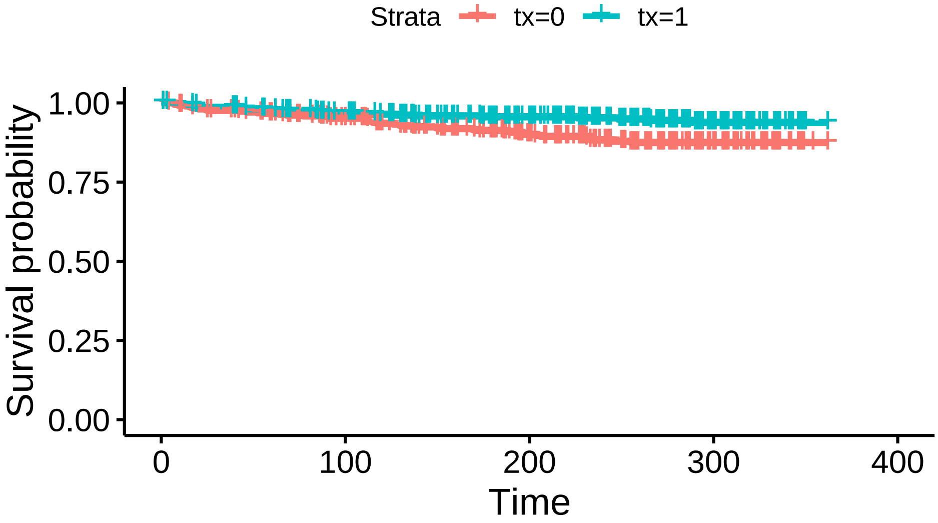


```
never<-sum(aids$ivdrug==1)
cur<-sum(aids$ivdrug==2)
prev<-sum(aids$ivdrug==3)
slices <- c(never,cur,prev)
lbls <- c("Never", "Currently", "Previously")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie3D(slices,labels=lbls,explode=0.1,col=c("turquoise","magenta","salmon"),cex.sub=0.5,
   main="IV Drug Use History ")
```

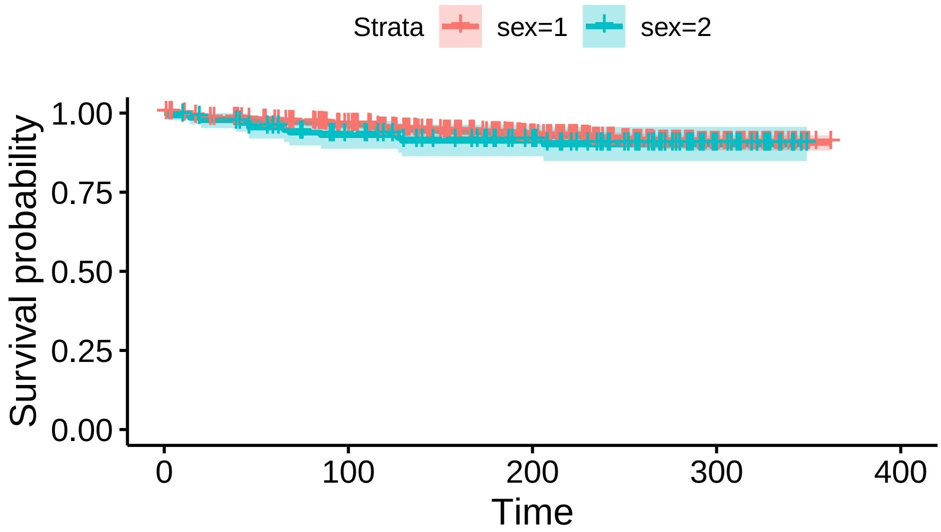


```
hist(aids$time)

###Data Plots

fit <- survfit(Surv(time,censor)~tx, data = aids)
ggsurvplot(fit,data = aids,conf.int = FALSE)
```

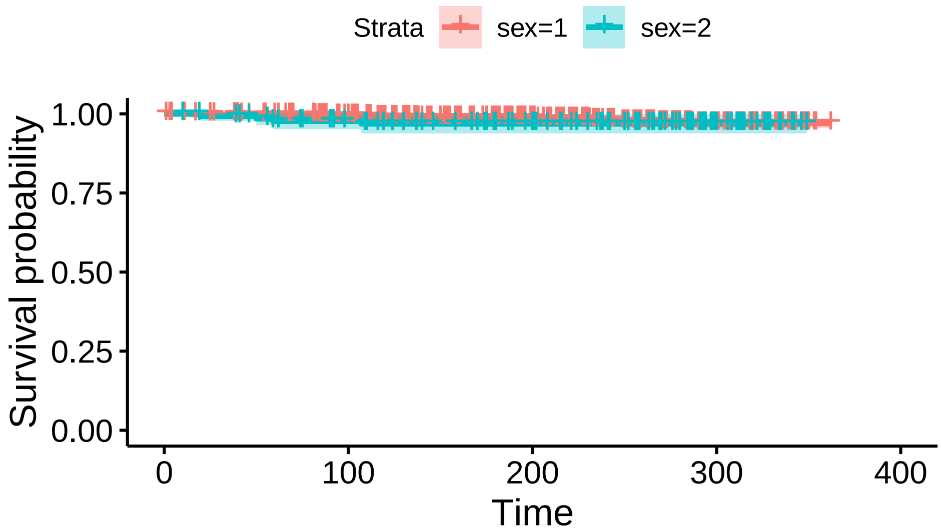




```
aids_fit_time <- survfit(Surv(time, censor) ~ sex, data=aids)
ggsurvplot(aids_fit_time, data=aids,  conf.int = TRUE)
```



```
aids_fit_time.d <- survfit(Surv(time_d, censor_d) ~ sex, data=aids)
ggsurvplot(aids_fit_time.d, data=aids,  conf.int = TRUE)
```



Survival Analysis

```
#mutation of age
aids <- read.csv( "http://pages.pomona.edu/~jsh04747/courses/math150/AIDSdata.csv")
aids <- aids %>%
  mutate(age = ifelse(age <= 20, "under20",
                             ifelse(age <=30, "20-30",
                                    ifelse(age <= 40, "30-40",
                                           ifelse(age <=50, "40-50",
                                                 ifelse(age <=60, "50-60",
                                                        ifelse(age <=70, "60-70", "over70"))))))) %>%
  mutate(age = factor(age,
                       levels = c("under20", "20-30", "30-40","40-50", "50-60","60-70","over70")),  sex
```

```
library(survival)
library (survminer)
library(ggplot2)
library(broom)

coxph(Surv(time_d,censor_d) ~ sex , data=aids)  %>% tidy()
```

```
## # A tibble: 1 x 7
##   term    estimate std.error statistic p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 sexmale    0.390     0.559     0.697   0.486   -0.706      1.49
```

```
coxph(Surv(time,censor) ~ sex, data=aids) %>% tidy()
```

```
## # A tibble: 1 x 7
##   term    estimate std.error statistic p.value conf.low conf.high
##   <chr>      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 sexmale    0.199     0.318     0.625   0.532   -0.424     0.821
```

```
coxph(Surv(time,censor) ~ age+ txgrp+ karnof, data=aids) %>% tidy()
```

```
## # A tibble: 8 x 7
##   term      estimate std.error statistic       p.value conf.low conf.high
##   <chr>        <dbl>     <dbl>     <dbl>         <dbl>    <dbl>     <dbl>
## 1 age20-30   -0.438      1.07    -0.409   0.682           -2.53      1.66
## 2 age30-40   -0.442      1.02    -0.434   0.665           -2.44      1.55
## 3 age40-50   -0.361      1.03    -0.352   0.725           -2.37      1.65
## 4 age50-60    0.460      1.04     0.442   0.659           -1.58      2.50
## 5 age60-70   -0.780      1.42    -0.551   0.582           -3.55      2.00
## 6 ageover70 -14.1     2688.      -0.00525 0.996          -Inf      Inf
## 7 txgrp      -0.844      0.257   -3.28    0.00103         -1.35    -0.340
## 8 karnof     -0.0814     0.0138  -5.89    0.00000000385   -0.109   -0.0543
```

```
cox.zph(coxph(Surv(time,censor) ~ age + txgrp+karnof, data=aids))
```

```
##                rho    chisq     p
## age20-30   0.09054 5.70e-01 0.450
## age30-40   0.19294 2.53e+00 0.112
## age40-50   0.14871 1.50e+00 0.220
## age50-60   0.19861 2.69e+00 0.101
## age60-70   0.16251 1.81e+00 0.179
## ageover70  0.16355 2.57e-07 1.000
## txgrp     -0.10779 8.34e-01 0.361
## karnof     0.00121 1.03e-04 0.992
## GLOBAL          NA 7.98e+00 0.435
```

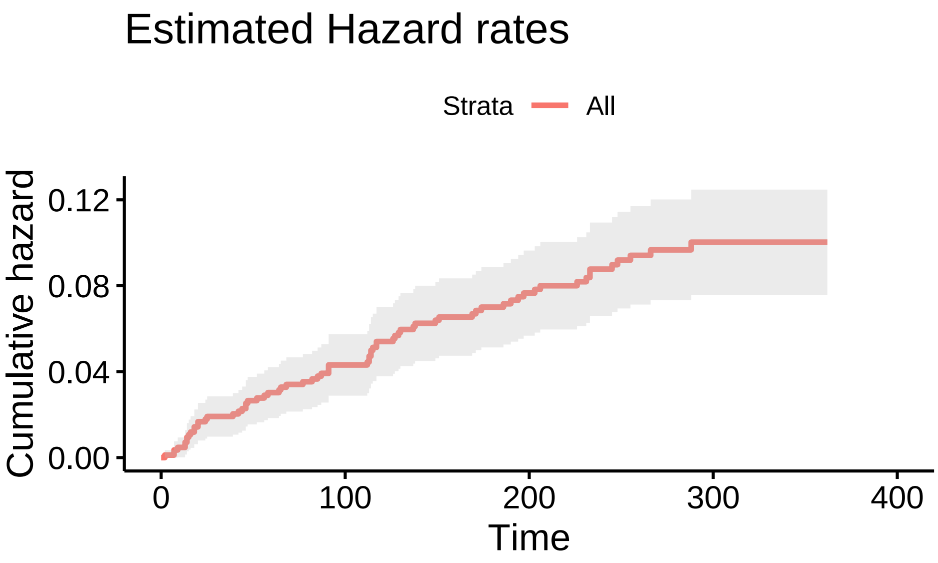
```
coxph(Surv(time,censor) ~ age *txgrp*karnof, data=aids) %>% tidy()
```

```
## # A tibble: 27 x 7
##    term           estimate std.error  statistic p.value conf.low conf.high
##    <chr>             <dbl>     <dbl>      <dbl>   <dbl>    <dbl>     <dbl>
##  1 age20-30         307.     138277.  0.00222     0.998     -Inf       Inf
##  2 age30-40         319.     138277.  0.00231     0.998     -Inf       Inf
##  3 age40-50         327.     138277.  0.00237     0.998     -Inf       Inf
##  4 age50-60         343.     138277.  0.00248     0.998     -Inf       Inf
##  5 age60-70         287.     176491.  0.00163     0.999     -Inf       Inf
##  6 ageover70         -1.66    29414. -0.0000565   1.000     -Inf       Inf
##  7 txgrp             150.      92392.  0.00163      0.999     -Inf       Inf
##  8 karnof              3.36     1424.  0.00236      0.998     -Inf       Inf
##  9 age20-30:txgrp   -144.      92392. -0.00156      0.999     -Inf       Inf
## 10 age30-40:txgrp   -146.      92392. -0.00158      0.999     -Inf       Inf
## # ... with 17 more rows
```

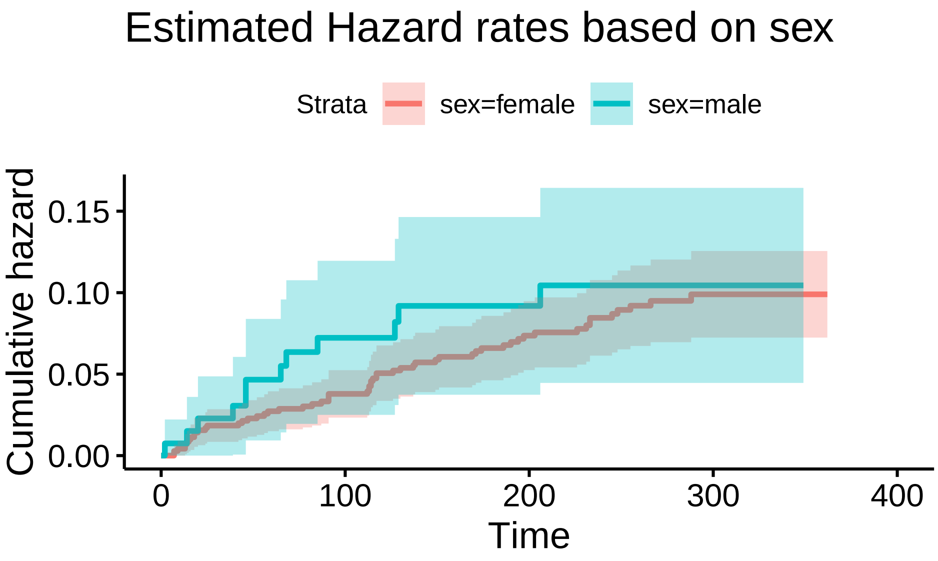
```r
cox.zph(coxph(Surv(time,censor) ~ age *txgrp*karnof, data=aids))
```

```
##                             rho    chisq     p
## age20-30                -0.1008 4.31e-08 1.000
## age30-40                -0.1583 3.15e-08 1.000
## age40-50                -0.0965 1.25e-08 1.000
## age50-60                -0.2071 6.53e-08 1.000
## age60-70                -0.2062 3.04e-08 1.000
## ageover70               -0.2493 7.81e-11 1.000
## txgrp                   -0.2032 2.68e-08 1.000
## karnof                  -0.1974 5.24e-08 1.000
## age20-30:txgrp           0.0921 2.14e-08 1.000
## age30-40:txgrp           0.1142 1.08e-08 1.000
## age40-50:txgrp           0.0826 5.64e-09 1.000
## age50-60:txgrp           0.1851 3.47e-08 1.000
## age60-70:txgrp           0.2102 2.15e-08 1.000
## ageover70:txgrp          0.1967 3.96e-11 1.000
## age20-30:karnof          0.0984 4.53e-08 1.000
## age30-40:karnof          0.1524 3.44e-08 1.000
## age40-50:karnof          0.0938 1.40e-08 1.000
## age50-60:karnof          0.2053 7.78e-08 1.000
## age60-70:karnof          0.1978 3.00e-08 1.000
## ageover70:karnof             NA      NaN   NaN
## txgrp:karnof             0.1996 2.81e-08 1.000
## age20-30:txgrp:karnof   -0.0910 2.15e-08 1.000
## age30-40:txgrp:karnof   -0.1020 9.71e-09 1.000
## age40-50:txgrp:karnof   -0.0823 6.23e-09 1.000
## age50-60:txgrp:karnof   -0.1796 3.72e-08 1.000
## age60-70:txgrp:karnof   -0.1981 1.98e-08 1.000
## ageover70:txgrp:karnof       NA      NaN   NaN
## GLOBAL                       NA 1.84e+01 0.891
```

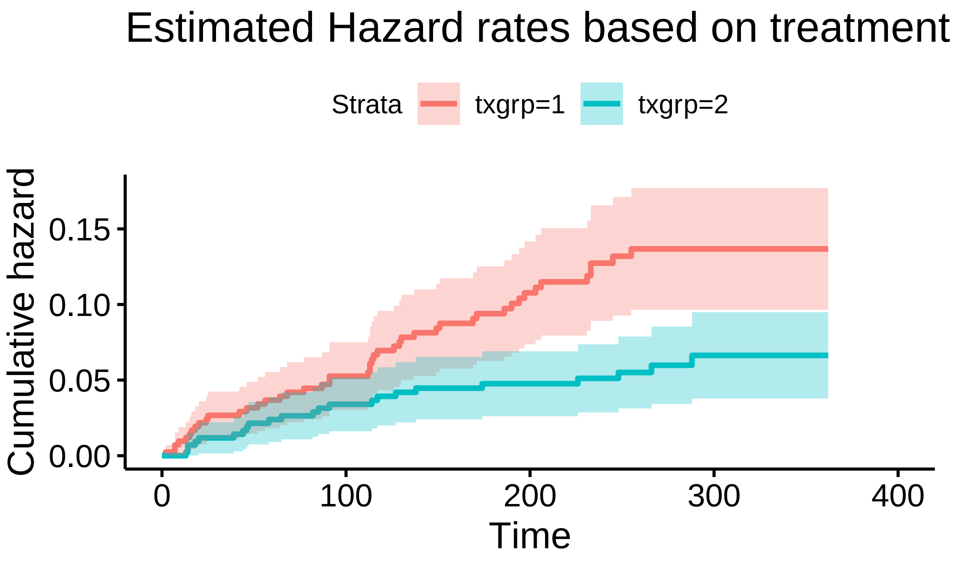
```r
ggsurvplot(survfit(Surv(time,censor) ~ 1, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates")
```



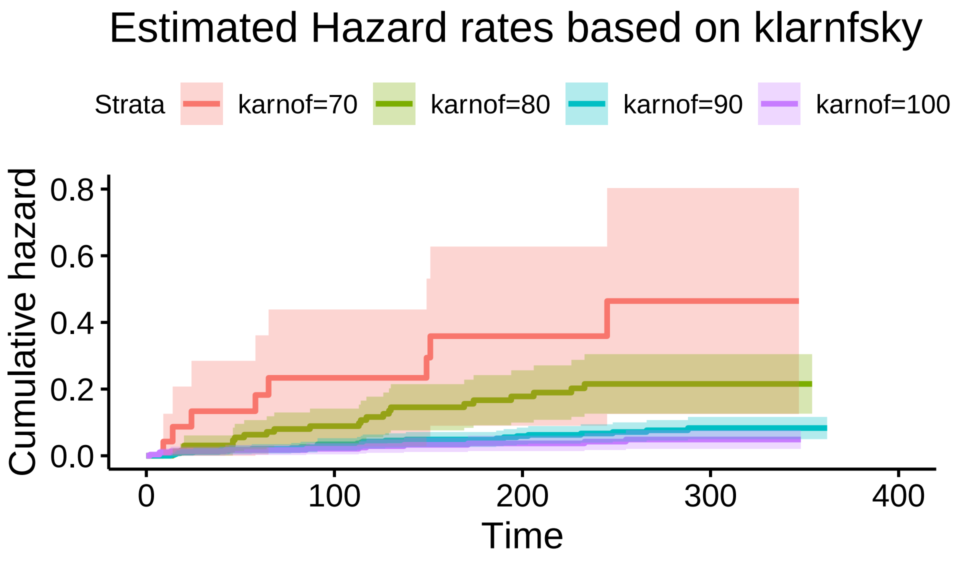
```r
ggsurvplot(survfit(Surv(time,censor) ~ sex, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on sex")
```



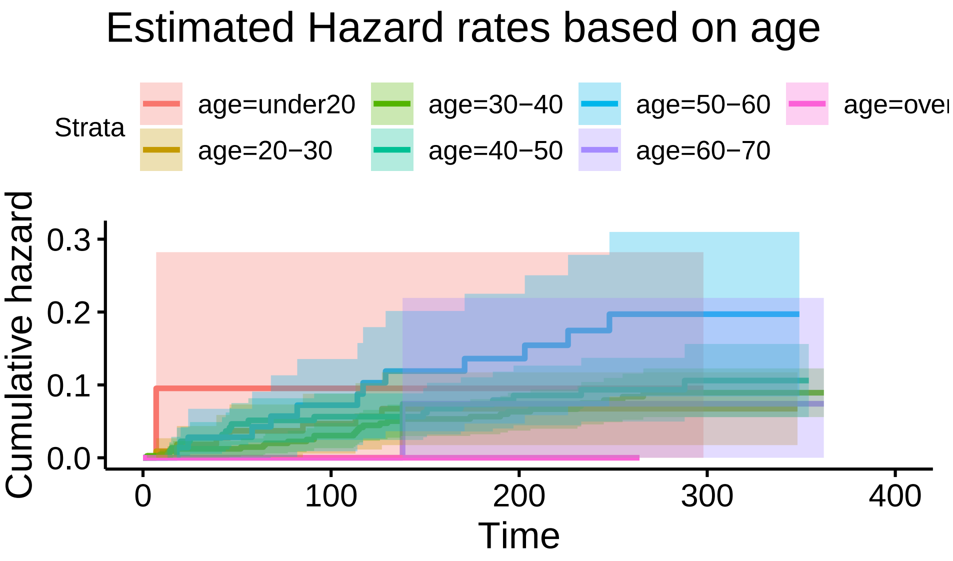
```r
ggsurvplot(survfit(Surv(time,censor) ~ txgrp, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on treatment grou
```



```
ggsurvplot(survfit(Surv(time,censor) ~ karnof, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on klarnfsky")
```



```
ggsurvplot(survfit(Surv(time,censor) ~ age, data=aids),
           censor=F, conf.int=T, fun="cumhaz") + ggtitle("Estimated Hazard rates based on age")
```



Juste's "Something New"

I will be analyzing the Weibull PH regression (parametric survival model).

1. What is goign on? What is the topic?

Weibull model is a parametric model,which provides a flexible way for the inclusion of covariates of the survival times. It has been previously determined that when the shape of parameter is known, the Weibull model shows better results than the Cox proportional hazards model, but when the shape parameter is unknown, the Cox proportional hazards model and the Weibull model give comparable results. ### 2. How it is relevant? How it relates to survival analysis/analysis at hand?

Fully parametric models have many advantages in analyzing survival data, they can be more convenient for representing complex data structures and processes. Studies have indicated that under certain situations when the shape of the survival time is determined, the parametric models are more powerful and efficient than Cox's regression model (ex. Kleinbaum D, Klein M. Survival analysis: a self-learning text. New York: Springer; 2005). Furthermore, if the only basic assumption of this model (proportional hazards) is not met, parametric models are suitable alternative models to be used instead of Cox's regression analysis.

3. Resources to learn about the topic.

I have been researching articles and scientific journals that provide insights into this model and comparisons between the Cox PH and teh parametric model. Sources include: a) https://krex.k-state.edu/dspace/bitstream/handle/2097/8787/AngelaCrumer2011.pdf b) http://nematilab.info/bmijc/assets/weibull_cox.pdf c) https://www.jstatsoft.org/article/view/v070i08

4. What will be challenging about learning something new?

Taking a completely new model of analyzing survival data is particulalrly difficult since the mathematical derivations and notations are also very varied from what we have seen in class. Although, I do remember some of the ideas behind parametric functions, their applications to statistical models are much more challenging than I have expected. Therefore, it will require me a lot of time and extensive research to be able to understand and learn how to apply this model to our data and other instances of survival analysis.

```
### some trials of applications of parametric functions in r
library(flexsurv)

flexsurvreg(Surv(time, censor) ~ age, data = aids,  dist = "weibull")
```

```
## Call:
## flexsurvreg(formula = Surv(time, censor) ~ age, data = aids,
##     dist = "weibull")
##
## Estimates:
##            data mean  est        L95%       U95%       se
## shape             NA   7.90e-01   6.30e-01   9.90e-01   9.10e-02
## scale             NA   4.17e+03   3.20e+02   5.43e+04   5.46e+03
## age20-30    1.30e-01   5.91e-01  -2.06e+00   3.25e+00   1.36e+00
## age30-40    4.89e-01   4.53e-01  -2.07e+00   2.98e+00   1.29e+00
## age40-50    2.64e-01   2.08e-01  -2.34e+00   2.75e+00   1.30e+00
## age50-60    8.46e-02  -5.81e-01  -3.17e+00   2.01e+00   1.32e+00
## age60-70    1.65e-02   6.27e-01  -2.88e+00   4.14e+00   1.79e+00
## ageover70   2.35e-03   1.88e+01  -8.97e+03   9.01e+03   4.59e+03
##            exp(est)   L95%       U95%
## shape             NA          NA          NA
## scale             NA          NA          NA
## age20-30    1.81e+00   1.27e-01   2.57e+01
## age30-40    1.57e+00   1.26e-01   1.97e+01
## age40-50    1.23e+00   9.65e-02   1.57e+01
## age50-60    5.60e-01   4.21e-02   7.45e+00
## age60-70    1.87e+00   5.59e-02   6.27e+01
## ageover70   1.51e+08   0.00e+00        Inf
##
## N = 851,  Events: 69,  Censored: 782
## Total time at risk: 197290
## Log-likelihood = -612.8653, df = 8
## AIC = 1241.731
```