Cox regression

A manually worked out, simple example: two groups

Load libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'dplyr' was built under R version 3.6.1

## -- Conflicts ------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(maxLik)
```

```
## Warning: package 'maxLik' was built under R version 3.6.1

## Loading required package: miscTools

## Warning: package 'miscTools' was built under R version 3.6.1

##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. Co
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum or
## https://r-forge.r-project.org/projects/maxlik/
```

```
library(survival)
```

Data definition

Lets enter the data in R:

```
dat <- data.frame(ratID = paste0("rat", 1:5),
                  time = c(55, 50, 70, 120, 110),
                  failure = c(0, 1, 1, 0, 1),
                  group = c(0, 1, 0, 1, 1))
```

Total number of failures D:

```r
sum(dat$failure)
```

```
## [1] 3
```

For convenience, rename 'group' to 'x':

```r
dat <- rename(dat, x = group)
dat
```

```
##   ratID time failure x
## 1  rat1   55       0 0
## 2  rat2   50       1 1
## 3  rat3   70       1 0
## 4  rat4  120       0 1
## 5  rat5  110       1 1
```

We also define an auxiliary data.frame containing events only:

```r
dat.events <- subset(dat, failure == 1)
```

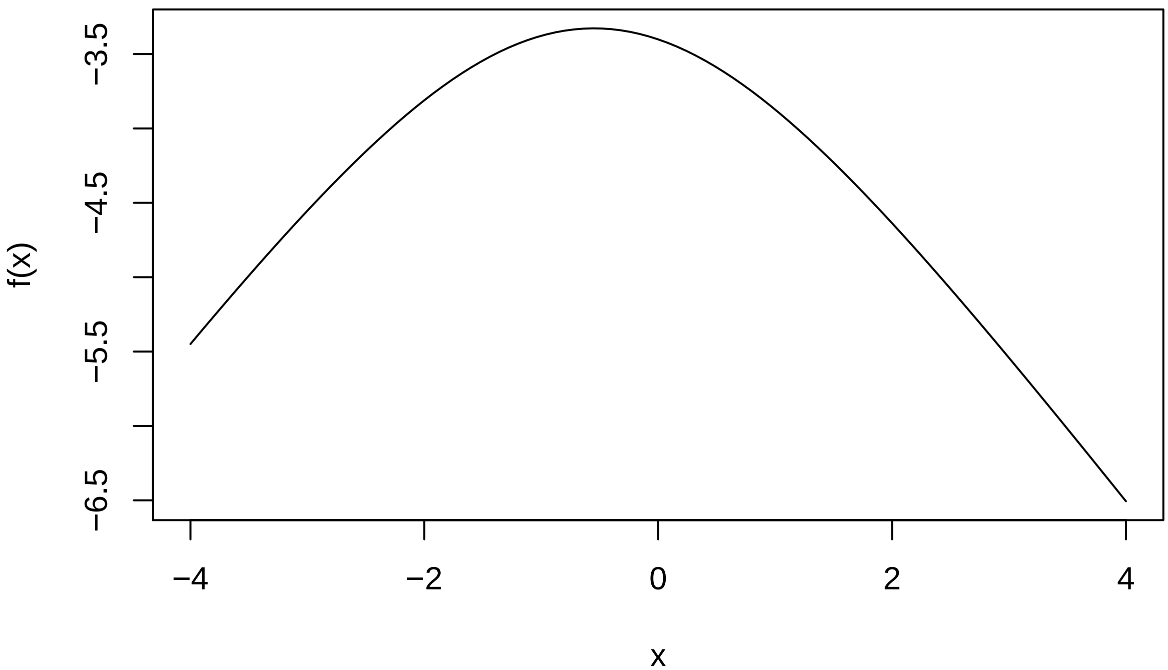
Partial log-likelihood function

Lets define the partial (log-)likelihood function

```r
pLogLik <- function(beta) {
  numerator <- with(dat.events, x * beta)
  denominator <- rep(NA_real_, length(numerator))
  for(j in seq_along(denominator)) {
    risk_set <- subset(dat, time >= dat.events[j, "time"])
    theta_j <- with(risk_set, exp(x * beta))# within the risk set, we compute the function for each rat
    denominator[j] <- log(sum(theta_j))
  }
  #with log, we only need to do sum, not product, to easier computation
  return(sum(numerator - denominator))
}
```

We can plot it:

```r
f <- Vectorize(pLogLik)
curve(f, from = -4, to = 4)
```



Maximum partial-Likelihood estimation

interpretation:

x_i

- 0: normal sleep pattern
- 1: sleep deprived

$h_i(t) = h_0(t)exp(x_iB)$ $\hat{B} = -0.55(SE = 1.4)$

Hazard ratio: (between 2 group)

$\frac{h_{SD}(t)=h_o*exp(1*-0.55)}{h_{NSD}(t)=h_o*exp(0*-0.55)} = exp(-0.55)$

```
fit.ML <- maxLik(pLogLik, start = c(beta = 0))
summary(fit.ML)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 2 iterations
## Return code 1: gradient close to zero
## Log-Likelihood: -3.327063
## 1  free parameters
## Estimates:
```

```
##      Estimate Std. error t value Pr(> t)
## beta  -0.5493     1.4179  -0.387   0.698
## --------------------------------------------
```

```
#hazard ratio
exp(-0.55)
```

```
## [1] 0.5769498
```

With the `coxph` function:

```
fit.cph <- coxph(Surv(time, failure) ~ x, data = dat)
summary(fit.cph)
```

```
## Call:
## coxph(formula = Surv(time, failure) ~ x, data = dat)
##
##   n= 5, number of events= 3
##
##      coef exp(coef) se(coef)      z Pr(>|z|)
## x -0.5493    0.5774   1.4179 -0.387    0.698
##
##   exp(coef) exp(-coef) lower .95 upper .95
## x    0.5774      1.732   0.03585     9.297
##
## Concordance= 0.5  (se = 0.202 )
## Likelihood ratio test= 0.15  on 1 df,   p=0.7
## Wald test            = 0.15  on 1 df,   p=0.7
## Score (logrank) test = 0.15  on 1 df,   p=0.7
```

We can reproduce the Likelihood-ratio test:

```
LRT <- 2 * (fit.ML$maximum - pLogLik(0))
data.frame(LRT = LRT,
           pvalue = pchisq(LRT, df = 1, lower.tail = FALSE))
```

```
##         LRT    pvalue
## 1 0.1482688 0.7001953
```

The Wald test is already in the `maxLik` summary output.

A manually worked out, simple example: one continuous covariate

```
dat <- data.frame(time = c(6, 7, 10, 15, 19, 25),
                  event = c(1, 0, 1, 1, 0, 1),
                  age = c(67, 62, 34, 41, 46, 28))
```

```
fit <- coxph(Surv(time, event) ~ age, data = dat)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ age, data = dat)
##
##   n= 6, number of events= 4
##
##         coef exp(coef) se(coef)    z Pr(>|z|)
## age 0.07606   1.07903  0.07316 1.04    0.298
##
##     exp(coef) exp(-coef) lower .95 upper .95
## age     1.079     0.9268    0.9349     1.245
##
## Concordance= 0.7  (se = 0.237 )
## Likelihood ratio test= 1.41  on 1 df,   p=0.2
## Wald test            = 1.08  on 1 df,   p=0.3
## Score (logrank) test = 1.33  on 1 df,   p=0.2
```

We might express age in decades:

```
dat <- mutate(dat, age_dec = age / 10)
summary(coxph(Surv(time, event) ~ age_dec, data = dat))
```

```
## Call:
## coxph(formula = Surv(time, event) ~ age_dec, data = dat)
##
##   n= 6, number of events= 4
##
##            coef exp(coef) se(coef)    z Pr(>|z|)
## age_dec 0.7606    2.1397   0.7316 1.04    0.298
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age_dec      2.14     0.4674      0.51     8.976
##
## Concordance= 0.7  (se = 0.237 )
## Likelihood ratio test= 1.41  on 1 df,   p=0.2
## Wald test            = 1.08  on 1 df,   p=0.3
## Score (logrank) test = 1.33  on 1 df,   p=0.2
```

Case study: the pharmacoSmoking dataset

Load the data

```
library(asaur)
dat <- pharmacoSmoking
head(dat)
```

```
##    id ttr relapse         grp age gender     race employment yearsSmoking
```

```
## 1  21 182       0   patchOnly  36   Male    white         ft           26
## 2 113  14       1   patchOnly  41   Male    white      other           27
## 3  39   5       1 combination  25 Female    white      other           12
## 4  80  16       1 combination  54   Male    white         ft           39
## 5  87   0       1 combination  45   Male    white      other           30
## 6  29 182       0 combination  43   Male hispanic         ft           30
##   levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
## 1        heavy     21-49     35-49             0              0
## 2        heavy     21-49     35-49             3             90
## 3        heavy     21-49     21-34             3             21
## 4        heavy       50+     50-64             0              0
## 5        heavy     21-49     35-49             0              0
## 6        heavy     21-49     35-49             2           1825
```

grp is not 0,1. R would transform it to be 0,1 (alph ordering combination:0, patchOnly:1), so that we see the risk H1>H0, the risk in pathOnly is higher so it means the time is shorter. P-value is small, so there's a significant difference.

```r
summary(coxph(Surv(ttr,relapse)~grp,data=dat))
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp, data = dat)
##
##   n= 125, number of events= 89
##
##                coef exp(coef) se(coef)   z Pr(>|z|)
## grppatchOnly 0.6050    1.8313   0.2161 2.8  0.00511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly     1.831     0.5461     1.199     2.797
##
## Concordance= 0.581  (se = 0.027 )
## Likelihood ratio test= 7.99  on 1 df,   p=0.005
## Wald test            = 7.84  on 1 df,   p=0.005
## Score (logrank) test = 8.07  on 1 df,   p=0.004
```

Fit the Cox model

```r
fit <- coxph(Surv(ttr, relapse) ~ grp + age + gender + priorAttempts, data = dat)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + age + gender + priorAttempts,
##     data = dat)
##
##   n= 125, number of events= 89
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## grppatchOnly   0.5656340  1.7605636  0.2181634  2.593  0.00952 **
```

```
## age           -0.0220948  0.9781475  0.0097572 -2.264  0.02355 *
## genderMale    -0.1215514  0.8855455  0.2334349 -0.521  0.60257
## priorAttempts  0.0002078  1.0002079  0.0010898  0.191  0.84876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly     1.7606     0.5680    1.1480     2.700
## age              0.9781     1.0223    0.9596     0.997
## genderMale       0.8855     1.1292    0.5604     1.399
## priorAttempts    1.0002     0.9998    0.9981     1.002
##
## Concordance= 0.623  (se = 0.031 )
## Likelihood ratio test= 14.14  on 4 df,   p=0.007
## Wald test            = 13.87  on 4 df,   p=0.008
## Score (logrank) test = 14.12  on 4 df,   p=0.007
```

We can change the contrasts as we see fit:

```
dat <- mutate(dat, grp = relevel(grp, ref = "patchOnly")) #change patchOnly to 0
fit <- update(fit)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + age + gender + priorAttempts,
##     data = dat)
##
##   n= 125, number of events= 89
##
##                     coef  exp(coef)   se(coef)      z Pr(>|z|)
## grpcombination -0.5656340  0.5679999  0.2181634 -2.593  0.00952 **
## age            -0.0220948  0.9781475  0.0097572 -2.264  0.02355 *
## genderMale     -0.1215514  0.8855455  0.2334349 -0.521  0.60257
## priorAttempts   0.0002078  1.0002079  0.0010898  0.191  0.84876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## grpcombination    0.5680     1.7606    0.3704    0.8711
## age               0.9781     1.0223    0.9596    0.9970
## genderMale        0.8855     1.1292    0.5604    1.3993
## priorAttempts     1.0002     0.9998    0.9981    1.0023
##
## Concordance= 0.623  (se = 0.031 )
## Likelihood ratio test= 14.14  on 4 df,   p=0.007
## Wald test            = 13.87  on 4 df,   p=0.008
## Score (logrank) test = 14.12  on 4 df,   p=0.007
```

encoding for categorical varianve more then 2 category (we can also change the reverance)

```
summary(coxph(Surv(ttr,relapse)~employment,data=dat)) #default fulltime is reference (decide alphabetic
```

```
## Call:
```

```
## coxph(formula = Surv(ttr, relapse) ~ employment, data = dat)
##
##   n= 125, number of events= 89
##
##                   coef exp(coef) se(coef)     z Pr(>|z|)
## employmentother 0.1982    1.2192   0.2371 0.836    0.403
## employmentpt    0.4500    1.5683   0.3229 1.394    0.163
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## employmentother     1.219     0.8202    0.7661     1.940
## employmentpt        1.568     0.6376    0.8328     2.953
##
## Concordance= 0.541  (se = 0.028 )
## Likelihood ratio test= 2.06  on 2 df,   p=0.4
## Wald test            = 2.17  on 2 df,   p=0.3
## Score (logrank) test = 2.2  on 2 df,   p=0.3
```

```r
#if we want to change the reference
dat1<-mutate(dat,employment =relevel(employment, ref="pt"))
summary(coxph(Surv(ttr,relapse)~employment,data=dat1))
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ employment, data = dat1)
##
##   n= 125, number of events= 89
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## employmentft    -0.4500    0.6376   0.3229 -1.394    0.163
## employmentother -0.2518    0.7774   0.3455 -0.729    0.466
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## employmentft       0.6376      1.568    0.3386     1.201
## employmentother    0.7774      1.286    0.3949     1.530
##
## Concordance= 0.541  (se = 0.028 )
## Likelihood ratio test= 2.06  on 2 df,   p=0.4
## Wald test            = 2.17  on 2 df,   p=0.3
## Score (logrank) test = 2.2  on 2 df,   p=0.3
```

Case study: the lung cancer dataset

Load the data

```r
library(survival)

dat <- lung
dat$delta<-dat$status-1
dat$S <-with(dat,Surv(time/365.25,delta))
#equivalant as Surv(dat$time,dat$delta)
#for S it store 2 value, but when we print it out when it's censoring it put a plus in the end
```

```r
#rescale time to year
head(dat)
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
##   delta          S
## 1     1  0.8377823
## 2     1  1.2457221
## 3     0 2.7652293+
## 4     1  0.5749487
## 5     1  2.4175222
## 6     0 2.7980835+
```

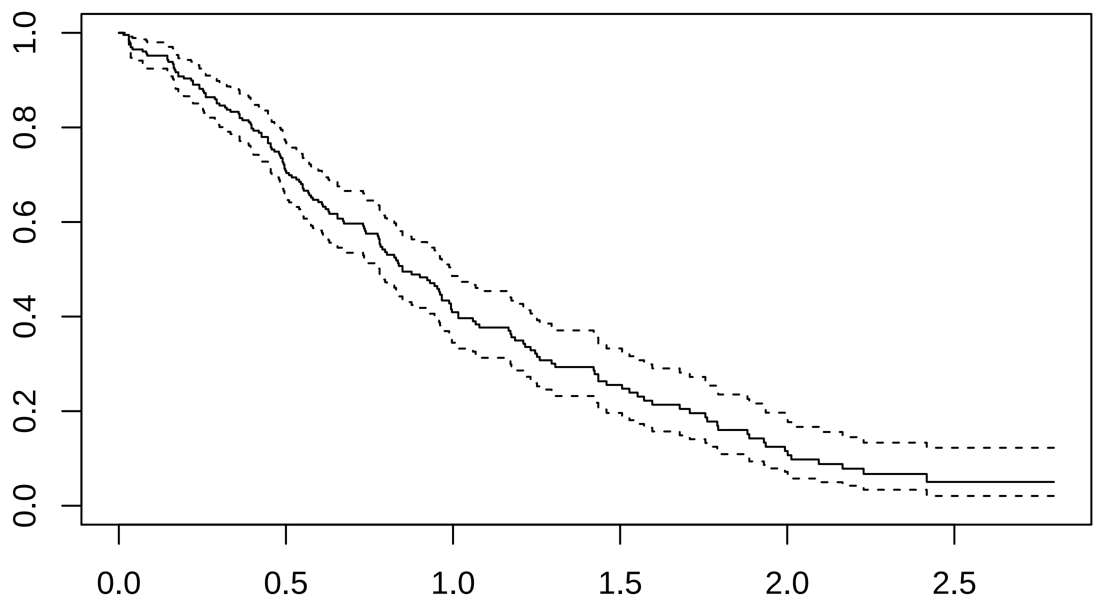
exercise:

Q1: median survival and confidence interval Q2: survival of men vs woman

- median survival with each group
- test & the diff
- HR?

Q3: is self-evaluate karno score equivilent to phisician's score?

```r
#Q1 medium survival
fit.KM <-survfit(S~1, data=dat) #survival curve go to 0, after 2.5 years almost every patients dead
plot(fit.KM)
```




```
fit.KM # we can get the median and confidence interval
```

```
## Call: survfit(formula = S ~ 1, data = dat)
##
##        n  events  median 0.95LCL 0.95UCL
## 228.000 165.000   0.849   0.780   0.994
```

```
#Q2
table(dat$sex, useNA = "always")# report also missing
```

```
##
##    1    2 <NA>
##  138   90    0
```

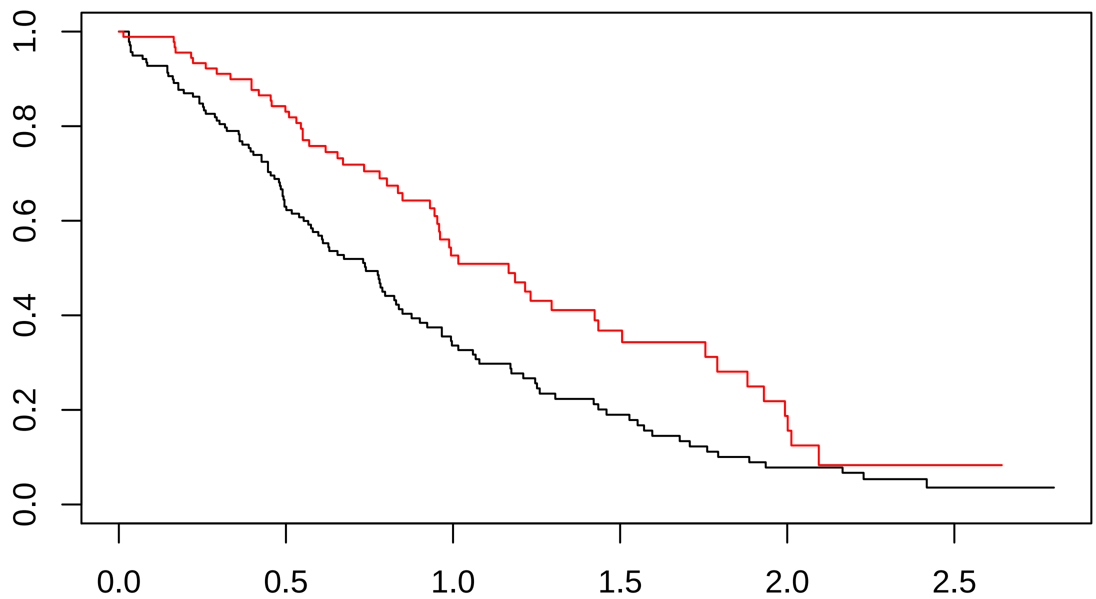
```
dat$sex <- factor(dat$sex,levels=1:2, labels=c("m","f"))
table(dat$sex, useNA = "always")
```

```
##
##    m    f <NA>
##  138   90    0
```

```
fit.KM <-survfit(S~sex, data=dat)
fit.KM
```

```
## Call: survfit(formula = S ~ sex, data = dat)
##
##          n events median 0.95LCL 0.95UCL
## sex=m 138    112  0.739   0.580   0.849
## sex=f  90     53  1.166   0.953   1.506
```

```
plot(fit.KM,col=1:2) #black:m red:female
```




```
#survival for man is worse then women
survdiff(S~sex, data=dat)
```

```
## Call:
## survdiff(formula = S ~ sex, data = dat)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=m 138      112     91.6      4.55      10.3
## sex=f  90       53     73.4      5.68      10.3
##
##  Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

```
#there's the significance difference

#HR?
summary(coxph(S~sex,data = dat))
```

```
## Call:
## coxph(formula = S ~ sex, data = dat)
##
##   n= 228, number of events= 165
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## sexf -0.5310    0.5880   0.1672 -3.176  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sexf     0.588      1.701    0.4237     0.816
##
## Concordance= 0.579  (se = 0.021 )
## Likelihood ratio test= 10.63  on 1 df,   p=0.001
## Wald test            = 10.09  on 1 df,   p=0.001
## Score (logrank) test = 10.33  on 1 df,   p=0.001
```

```
#risk for man is higher, it's significantly different
#confirm the result of the survival test
```

Q3:

```
summary(dat)# supposely they should be similar, self one has larger range
```

```
##       inst             time            status           age         sex
##  Min.   : 1.00   Min.   :   5.0   Min.   :1.000   Min.   :39.00   m:138
##  1st Qu.: 3.00   1st Qu.: 166.8   1st Qu.:1.000   1st Qu.:56.00   f: 90
##  Median :11.00   Median : 255.5   Median :2.000   Median :63.00
##  Mean   :11.09   Mean   : 305.2   Mean   :1.724   Mean   :62.45
##  3rd Qu.:16.00   3rd Qu.: 396.5   3rd Qu.:2.000   3rd Qu.:69.00
##  Max.   :33.00   Max.   :1022.0   Max.   :2.000   Max.   :82.00
##  NA's   :1
##     ph.ecog          ph.karno        pat.karno         meal.cal
##  Min.   :0.0000   Min.   : 50.00   Min.   : 30.00   Min.   :  96.0
##  1st Qu.:0.0000   1st Qu.: 75.00   1st Qu.: 70.00   1st Qu.: 635.0
##  Median :1.0000   Median : 80.00   Median : 80.00   Median : 975.0
##  Mean   :0.9515   Mean   : 81.94   Mean   : 79.96   Mean   : 928.8
##  3rd Qu.:1.0000   3rd Qu.: 90.00   3rd Qu.: 90.00   3rd Qu.:1150.0
##  Max.   :3.0000   Max.   :100.00   Max.   :100.00   Max.   :2600.0
##  NA's   :1        NA's   :1        NA's   :3        NA's   :47
##     wt.loss             delta
##  Min.   :-24.000   Min.   :0.0000
##  1st Qu.:  0.000   1st Qu.:0.0000
##  Median :  7.000   Median :1.0000
##  Mean   :  9.832   Mean   :0.7237
##  3rd Qu.: 15.750   3rd Qu.:1.0000
##  Max.   : 68.000   Max.   :1.0000
##  NA's   :14
##        S.time              S.status
##  Min.   :0.0136893    Min.   :0.0000000
##  1st Qu.:0.4565366    1st Qu.:0.0000000
##  Median :0.6995209    Median :1.0000000
```

```
##  Mean   :0.8356809    Mean   :0.7236842
##  3rd Qu.:1.0855578    3rd Qu.:1.0000000
##  Max.   :2.7980835    Max.   :1.0000000
##
```

Doctor HR: 0.9837 (0.9725,0.995) Patient HR: 0.980 (0.970,0.991) so we can kind of assume that they are redunctant

- it's better to reverse the ratio. because saying it 2 times more is better then saying it's 0.5 times: 1/0.983=1.017
- it's not significant 1.017 to explain so we can also do some transformation
- we can now expan now the observation is 1.17 per 10 units **decrese**(we use the exp(-coef)as reference) of the score

```r
#summary(coxph(S~ph.karno,data=dat))
#summary(coxph(S~pat.karno,data=dat))
summary(coxph(S~I(pat.karno/10),data=dat))
```

```
## Call:
## coxph(formula = S ~ I(pat.karno/10), data = dat)
##
##   n= 225, number of events= 162
##    (3 observations deleted due to missingness)
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## I(pat.karno/10) -0.19850   0.81996  0.05467 -3.631 0.000282 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## I(pat.karno/10)      0.82       1.22    0.7366    0.9127
##
## Concordance= 0.607  (se = 0.025 )
## Likelihood ratio test= 12.47  on 1 df,   p=4e-04
## Wald test            = 13.18  on 1 df,   p=3e-04
## Score (logrank) test = 13.23  on 1 df,   p=3e-04
```

what if we try another model with both?

```r
summary(coxph(S~I(pat.karno/10)+I(ph.karno/10),data=dat))
```

```
## Call:
## coxph(formula = S ~ I(pat.karno/10) + I(ph.karno/10), data = dat)
##
##   n= 224, number of events= 161
##    (4 observations deleted due to missingness)
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## I(pat.karno/10) -0.16275   0.84980  0.06372 -2.554   0.0107 *
## I(ph.karno/10)  -0.07404   0.92863  0.06959 -1.064   0.2873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                   exp(coef) exp(-coef) lower .95 upper .95
## I(pat.karno/10)      0.8498      1.177    0.7500    0.9629
## I(ph.karno/10)       0.9286      1.077    0.8102    1.0643
##
## Concordance= 0.616  (se = 0.025 )
## Likelihood ratio test= 13.3  on 2 df,   p=0.001
## Wald test            = 14.01  on 2 df,   p=9e-04
## Score (logrank) test = 14.13  on 2 df,   p=9e-04
```

```
#taking each by each, the effect is very close
#but if we put them together, the doctor one drop
#the model is telling us it's redunctant to one and another
#the pvalue if we fix one, the another would not have much impact on the model
```

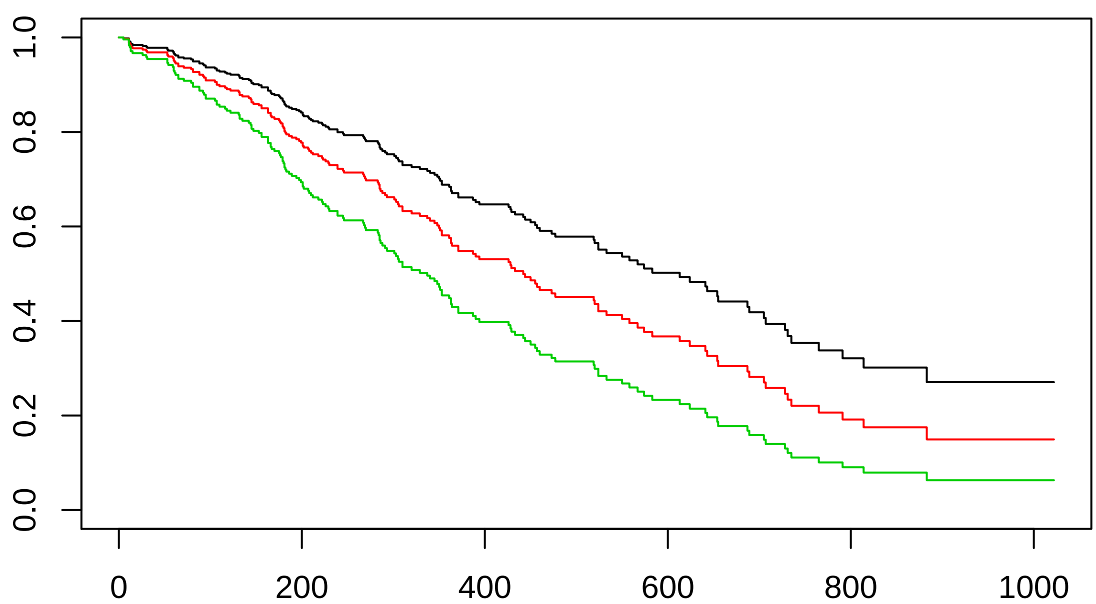
Cox regression: predictions

gnerally we don't use it in practice

```
fit.cph <- coxph(Surv(time, status) ~ age, data = dat)

pred.cph <- survfit(fit.cph, newdata = data.frame(age = c(20,40,60)))

plot(pred.cph, col = 1:3)
```



```r
#one cure for each indivisual datadrame
print(pred.cph)# base on cox model fix, we have a median of the point estimator
```

```
## Call: survfit(formula = fit.cph, newdata = data.frame(age = c(20, 40, 
##     60)))
## 
##     n events median 0.95LCL 0.95UCL
## 1 228    165    613     363      NA
## 2 228    165    442     329     705
## 3 228    165    337     288     371
```