Cox Model Building and Diagnostics

Model building

Load the data

```
library(tidyverse)
```

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

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

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

```
library(survival)
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:rpart':
##
##     solder
```

```
library(asaur)

dat <- pharmacoSmoking
```

The 4 candidate models

MA and MB arenested into MC, but not into eachother

```
M0 <- coxph(Surv(ttr, relapse) ~ 1, data = dat)
MA <- coxph(Surv(ttr, relapse) ~ ageGroup4, data = dat)
MB <- coxph(Surv(ttr, relapse) ~ employment, data = dat)
MC <- coxph(Surv(ttr, relapse) ~ ageGroup4 + employment, data = dat)
```

```
d <- mutate(dat, employment = ifelse(employment == "ft", "ft", "other"))
m_addicitve <- coxph(Surv(ttr, relapse) ~ grp + employment, data = d)
m_int <- coxph(Surv(ttr, relapse) ~ grp + employment + grp:employment, data = d)
anova(m_addicitve, m_int)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(ttr, relapse)
##  Model 1: ~ grp + employment
##  Model 2: ~ grp + employment + grp:employment
##    loglik  Chisq Df P(>|Chi|)
## 1 -380.81
## 2 -379.66 2.3054  1    0.1289
```

```
d$race <- relevel(d$race, ref = "other")
fit <- coxph(Surv(ttr, relapse) ~ grp + employment + gender + race + age, data = d)
summary(fit)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + employment + gender +
##     race + age, data = d)
##
##   n= 125, number of events= 89
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly     0.60734   1.83555  0.21816  2.784 0.005370 **
## employmentother  0.73101   2.07717  0.24162  3.025 0.002483 **
## genderMale      -0.04858   0.95258  0.23698 -0.205 0.837567
## raceblack        1.15692   3.18012  1.02155  1.133 0.257417
## racehispanic     0.60133   1.82455  1.09889  0.547 0.584228
## racewhite        0.89984   2.45922  1.01367  0.888 0.374696
## age             -0.03578   0.96485  0.01071 -3.342 0.000831 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.8355     0.5448    1.1969    2.8149
## employmentother    2.0772     0.4814    1.2936    3.3354
## genderMale         0.9526     1.0498    0.5987    1.5157
## raceblack          3.1801     0.3145    0.4294   23.5497
## racehispanic       1.8245     0.5481    0.2117   15.7227
## racewhite          2.4592     0.4066    0.3373   17.9322
## age                0.9649     1.0364    0.9448    0.9853
##
## Concordance= 0.652  (se = 0.03 )
## Rsquare= 0.182   (max possible= 0.998 )
## Likelihood ratio test= 25.14  on 7 df,   p=7e-04
## Wald test            = 24.53  on 7 df,   p=9e-04
## Score (logrank) test = 25.27  on 7 df,   p=7e-04
```

```
fit_NoRaceNoGnd <- coxph(Surv(ttr, relapse) ~ grp + employment + age, data = d)
anova(fit_NoRaceNoGnd, fit)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(ttr, relapse)
##  Model 1: ~ grp + employment + age
##  Model 2: ~ grp + employment + gender + race + age
##    loglik  Chisq Df P(>|Chi|)
## 1 -375.15
## 2 -373.58 3.1336  4    0.5357
```

Comparing nested models: LRT

```
anova(MA, MC) #anova(simple model, complex, model)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(ttr, relapse)
##  Model 1: ~ ageGroup4
##  Model 2: ~ ageGroup4 + employment
##    loglik  Chisq Df P(>|Chi|)
## 1 -380.04
## 2 -377.76 4.5666  2    0.1019
```

Do not reject null. The association with employment is not significantly different from 0. So, model A may be good enough

Comparing non-nested models: AIC

```
fits <- list(MA = MA, MB = MB, MC = MC)
sapply(fits, AIC)
```

```
##       MA       MB       MC
## 766.0860 774.2464 765.5194
```

```
list(MA, MB, MC)
```

```
## [[1]]
## Call:
## coxph(formula = Surv(ttr, relapse) ~ ageGroup4, data = dat)
##
##                     coef exp(coef) se(coef)      z      p
## ageGroup435-49  0.0293    1.0297   0.3093  0.095 0.9245
## ageGroup450-64 -0.7914    0.4532   0.3361 -2.355 0.0185
## ageGroup465+   -0.3173    0.7281   0.4435 -0.715 0.4744
##
## Likelihood ratio test=12.22  on 3 df, p=0.006664
## n= 125, number of events= 89
##
## [[2]]
## Call:
## coxph(formula = Surv(ttr, relapse) ~ employment, data = dat)
##
##                     coef exp(coef) se(coef)     z     p
## employmentother 0.1982    1.2192   0.2371 0.836 0.403
## employmentpt    0.4500    1.5683   0.3229 1.394 0.163
##
## Likelihood ratio test=2.06  on 2 df, p=0.357
## n= 125, number of events= 89
##
## [[3]]
## Call:
## coxph(formula = Surv(ttr, relapse) ~ ageGroup4 + employment,
##     data = dat)
##
##                    coef exp(coef) se(coef)      z       p
## ageGroup435-49  -0.1299    0.8782   0.3213 -0.404 0.68594
## ageGroup450-64  -1.0239    0.3592   0.3585 -2.856 0.00429
## ageGroup465+    -0.7825    0.4573   0.5046 -1.551 0.12102
## employmentother  0.5257    1.6917   0.2748  1.913 0.05577
## employmentpt     0.5001    1.6489   0.3315  1.508 0.13143
##
## Likelihood ratio test=16.79  on 5 df, p=0.004922
## n= 125, number of events= 89
```

```
sapply(list(ageOnly = MA, emplOnly = MB, full = MC), AIC)
```

```
##  ageOnly emplOnly     full
## 766.0860 774.2464 765.5194
```

Automatic model selection based on AIC

Remove parameter step by step

```
Mfull <- coxph(Surv(ttr, relapse) ~ grp + gender + race +
                 employment + yearsSmoking + levelSmoking +
                 ageGroup4 + priorAttempts + longestNoSmoke,
               data = dat)
```

```
MAIC <- step(Mfull)
```

```
## Start:  AIC=770.2
## Surv(ttr, relapse) ~ grp + gender + race + employment + yearsSmoking +
##     levelSmoking + ageGroup4 + priorAttempts + longestNoSmoke
##
##                  Df    AIC
## - race            3 766.98
## - yearsSmoking    1 768.20
## - gender          1 768.20
## - priorAttempts   1 768.24
## - levelSmoking    1 768.47
## - longestNoSmoke  1 769.04
## <none>              770.20
## - employment      2 772.45
## - ageGroup4       3 774.11
## - grp             1 776.80
##
## Step:  AIC=766.98
## Surv(ttr, relapse) ~ grp + gender + employment + yearsSmoking +
##     levelSmoking + ageGroup4 + priorAttempts + longestNoSmoke
##
##                  Df    AIC
## - levelSmoking    1 764.98
## - gender          1 765.00
## - priorAttempts   1 765.01
## - yearsSmoking    1 765.04
## - longestNoSmoke  1 766.29
## <none>              766.98
## - employment      2 768.37
## - ageGroup4       3 770.16
## - grp             1 773.88
##
## Step:  AIC=764.98
## Surv(ttr, relapse) ~ grp + gender + employment + yearsSmoking +
##     ageGroup4 + priorAttempts + longestNoSmoke
##
##                  Df    AIC
## - gender          1 763.00
## - priorAttempts   1 763.01
## - yearsSmoking    1 763.06
## - longestNoSmoke  1 764.29
## <none>              764.98
## - employment      2 766.37
## - ageGroup4       3 768.18
## - grp             1 771.88
##
## Step:  AIC=763
## Surv(ttr, relapse) ~ grp + employment + yearsSmoking + ageGroup4 +
##     priorAttempts + longestNoSmoke
##
##                  Df    AIC
## - priorAttempts   1 761.02
## - yearsSmoking    1 761.08
## - longestNoSmoke  1 762.31
## <none>              763.00
## - employment      2 764.42
## - ageGroup4       3 766.32
## - grp             1 769.91
##
## Step:  AIC=761.02
## Surv(ttr, relapse) ~ grp + employment + yearsSmoking + ageGroup4 +
##     longestNoSmoke
##
##                  Df    AIC
## - yearsSmoking    1 759.10
## - longestNoSmoke  1 760.34
## <none>              761.02
## - employment      2 762.42
## - ageGroup4       3 764.50
## - grp             1 767.93
##
## Step:  AIC=759.1
## Surv(ttr, relapse) ~ grp + employment + ageGroup4 + longestNoSmoke
##
##                  Df    AIC
## - longestNoSmoke  1 758.42
## <none>              759.10
## - employment      2 760.42
## - grp             1 765.94
## - ageGroup4       3 766.90
##
## Step:  AIC=758.42
## Surv(ttr, relapse) ~ grp + employment + ageGroup4
##
##              Df    AIC
## <none>          758.42
## - employment  2 760.31
## - grp         1 765.52
## - ageGroup4   3 767.24
```

```r
summary(MAIC)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + employment + ageGroup4,
##     data = dat)
##
##   n= 125, number of events= 89
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly     0.6564    1.9278   0.2198  2.986  0.00283 **
## employmentother  0.6231    1.8648   0.2764  2.254  0.02418 *
## employmentpt     0.5214    1.6844   0.3320  1.570  0.11631
## ageGroup435-49  -0.1119    0.8942   0.3216 -0.348  0.72792
## ageGroup450-64  -1.0233    0.3594   0.3597 -2.845  0.00444 **
## ageGroup465+    -0.7071    0.4931   0.5017 -1.410  0.15868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.9278     0.5187    1.2529    2.9661
## employmentother    1.8648     0.5363    1.0848    3.2057
## employmentpt       1.6844     0.5937    0.8787    3.2289
## ageGroup435-49     0.8942     1.1184    0.4761    1.6793
## ageGroup450-64     0.3594     2.7825    0.1776    0.7273
## ageGroup465+       0.4931     2.0281    0.1845    1.3180
##
## Concordance= 0.647  (se = 0.033 )
## Rsquare= 0.187   (max possible= 0.998 )
## Likelihood ratio test= 25.89  on 6 df,   p=2e-04
## Wald test            = 24.59  on 6 df,   p=4e-04
## Score (logrank) test = 25.54  on 6 df,   p=3e-04
```

Predictive power: concordance index

The higher the better

```r
summary(MA)
```

```
## Call:
```

```
## coxph(formula = Surv(ttr, relapse) ~ ageGroup4, data = dat)
##
##   n= 125, number of events= 89
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## ageGroup435-49  0.0293    1.0297   0.3093  0.095   0.9245
## ageGroup450-64 -0.7914    0.4532   0.3361 -2.355   0.0185 *
## ageGroup465+   -0.3173    0.7281   0.4435 -0.715   0.4744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## ageGroup435-49    1.0297     0.9711    0.5616    1.8880
## ageGroup450-64    0.4532     2.2066    0.2345    0.8757
## ageGroup465+      0.7281     1.3734    0.3053    1.7367
##
## Concordance= 0.593  (se = 0.032 )
## Rsquare= 0.093   (max possible= 0.998 )
## Likelihood ratio test= 12.22  on 3 df,   p=0.007
## Wald test            = 11.36  on 3 df,   p=0.01
## Score (logrank) test = 11.93  on 3 df,   p=0.008
```

```
summary(MAIC)
```

```
## Call:
## coxph(formula = Surv(ttr, relapse) ~ grp + employment + ageGroup4,
##     data = dat)
##
##   n= 125, number of events= 89
##
##                    coef exp(coef) se(coef)      z Pr(>|z|)
## grppatchOnly     0.6564    1.9278   0.2198  2.986  0.00283 **
## employmentother  0.6231    1.8648   0.2764  2.254  0.02418 *
## employmentpt     0.5214    1.6844   0.3320  1.570  0.11631
## ageGroup435-49  -0.1119    0.8942   0.3216 -0.348  0.72792
## ageGroup450-64  -1.0233    0.3594   0.3597 -2.845  0.00444 **
## ageGroup465+    -0.7071    0.4931   0.5017 -1.410  0.15868
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## grppatchOnly       1.9278     0.5187    1.2529    2.9661
## employmentother    1.8648     0.5363    1.0848    3.2057
## employmentpt       1.6844     0.5937    0.8787    3.2289
## ageGroup435-49     0.8942     1.1184    0.4761    1.6793
## ageGroup450-64     0.3594     2.7825    0.1776    0.7273
## ageGroup465+       0.4931     2.0281    0.1845    1.3180
##
## Concordance= 0.647  (se = 0.033 )
## Rsquare= 0.187   (max possible= 0.998 )
## Likelihood ratio test= 25.89  on 6 df,   p=2e-04
## Wald test            = 24.59  on 6 df,   p=4e-04
## Score (logrank) test = 25.54  on 6 df,   p=3e-04
```

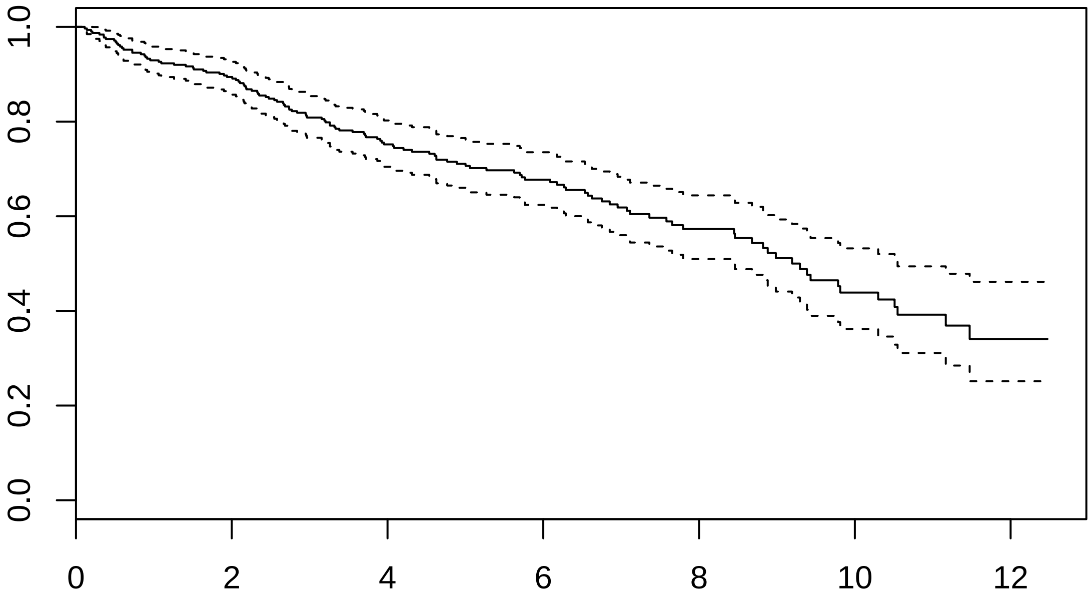
Predictive power: AUC

The higher the better

```
library(survivalROC)
data(mayo)
head(mayo)
```

```
##   time censor mayoscore5 mayoscore4
## 1   41      1  11.251850  10.629450
## 2  179      1  10.136070  10.185220
## 3  334      1  10.095740   9.422995
## 4  400      1  10.189150   9.567799
## 5  130      1   9.770148   9.039419
## 6  223      1   9.226429   9.033388
```

```
plot(survfit(Surv(time / 365.25, censor) ~ 1, data = mayo))
```



Pick a time point (365.25 * 5)

```
ROC.4 <- survivalROC(Stime = mayo$time,
                     status = mayo$censor,
                     marker = mayo$mayoscore4,
                     predict.time = 365.25 * 5,
                     method="KM")
```

```
ROC.5 <- survivalROC(Stime = mayo$time,
                     status = mayo$censor,
                     marker = mayo$mayoscore5,
                     predict.time = 365.25 * 5,
                     method = "KM")
```

```
ROC <- list(mayo4 = ROC.4, mayo5 = ROC.5)
map_dbl(ROC, "AUC")
```

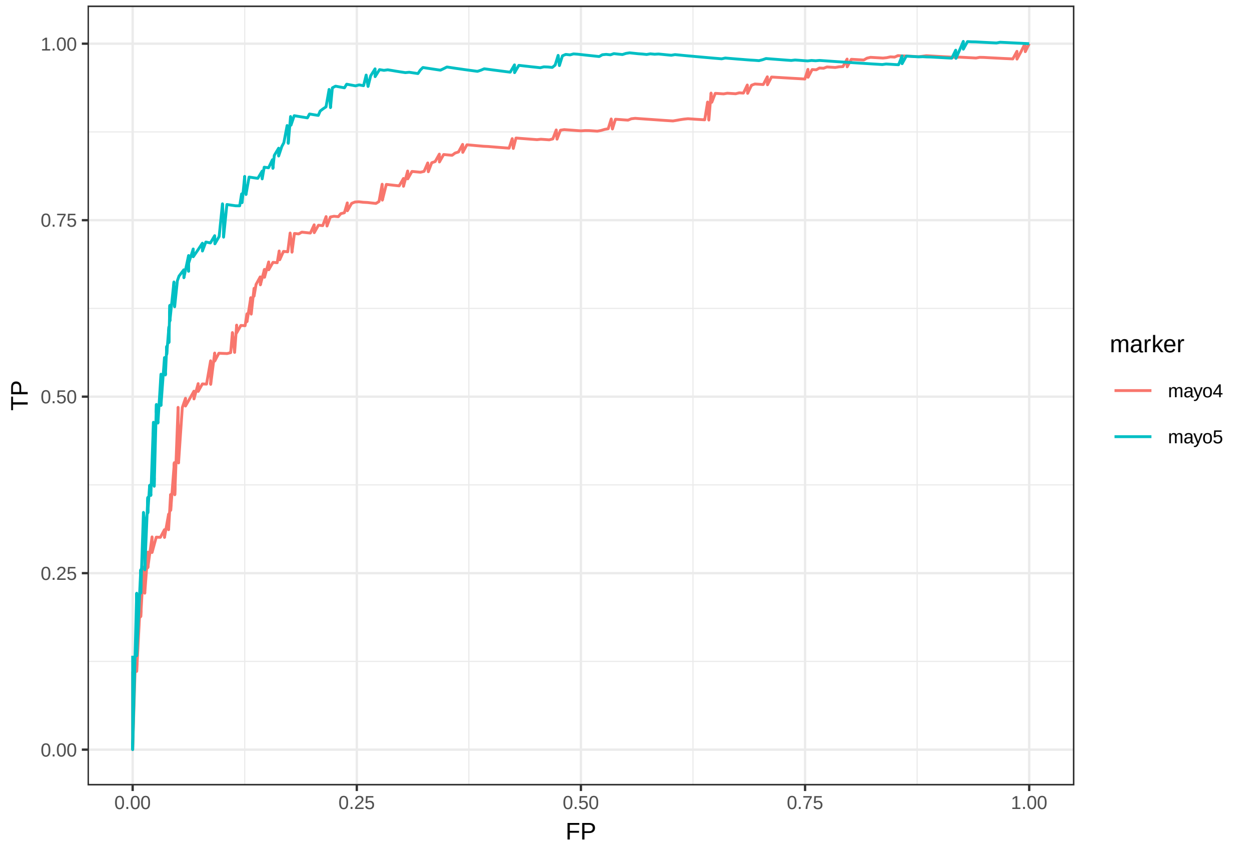
```
##     mayo4     mayo5
## 0.8257006 0.9182824
```

```
dfl <- map(ROC, ~ with(., tibble(cutoff = cut.values, FP, TP)))
for(nm in names(dfl)) {
  dfl[[ nm ]]$marker <- nm
}
dat <- do.call(rbind, dfl)
```

```
dat
```

```
## # A tibble: 626 x 4
##     cutoff    FP    TP marker
##  *   <dbl> <dbl> <dbl> <chr>
##  1 -Inf    1     1     mayo4
##  2    4.58 0.995 1.00  mayo4
##  3    4.90 0.996 0.989 mayo4
##  4    4.93 0.991 0.989 mayo4
##  5    4.93 0.986 0.989 mayo4
##  6    4.95 0.986 0.978 mayo4
##  7    4.97 0.982 0.978 mayo4
##  8    4.98 0.977 0.979 mayo4
##  9    5.06 0.972 0.979 mayo4
## 10    5.09 0.968 0.979 mayo4
## # ... with 616 more rows
```

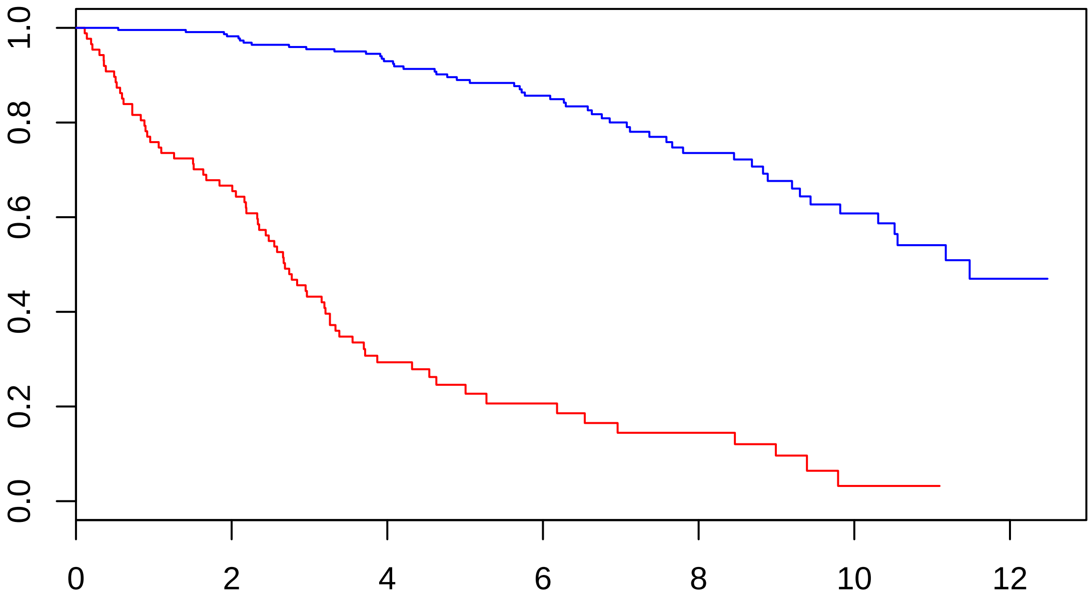
```
ggplot(dat, aes(FP, TP, color = marker)) +
  geom_line() +
  theme_bw(base_size = 9)
```



```
cutoff <- min(filter(dat, marker == "mayo5", FP <= 0.1)$cutoff)
```

```
mayo$prediction <-
  ifelse(mayo$mayoscore5 <= cutoff,
         "low_risk", "high_risk")
```

```
plot(survfit(Surv(time/365, censor) ~ prediction, data = mayo),
     col = c("red", "blue"))
```


1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12