Assignment 13; STAT 689

Philip Anderson; panders2@tamu.edu

4/11/2018

```
rm(list=ls())
library("tidyverse")
library("mgcv")
```

```
colon <- read.csv('/Users/panders2/Documents/schools/tamu/stat_689/homework/semiparametric-regression/m
str(colon)
```

```
## 'data.frame':    6696 obs. of  5 variables:
##  $ colon_cc : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ nonsmoker: int  1 1 0 1 1 1 1 1 1 0 ...
##  $ age      : num  63.3 70.2 58 66.9 68.9 ...
##  $ bmi      : num  25.9 26.6 29.6 30.3 25 ...
##  $ personyrs: num  5.16 4.51 1.71 2.13 2.12 ...
```

Question 1

What percentage of the sample died?

```
print("Sample death contingency table: ")
```

```
## [1] "Sample death contingency table: "
```

```
table(colon$colon_cc)
```

```
##
##    0    1
## 3348 3348
```

```
cat("\nSample death proportions: \n")
```

```
##
## Sample death proportions:
```

```
table(colon$colon_cc) / nrow(colon)
```

```
##
##   0   1
## 0.5 0.5
```

50% of the sample died from colon cancer.

Question 2

Run a Cox regression with age and BMI entering linearly; show output and code.

```
# mgcv expects the weights parameter to hold the censoring info.  0 for censored, 1 for event
# colon_cc = 0 in case of survival (censor), 1 if event occurs
cox_mod <- mgcv::gam(personyrs ~ age + bmi
                     , weights=colon_cc
                     , data=colon
                     , family="cox.ph"
                     )
mgcv::summary.gam(cox_mod)
```

```
##
## Family: Cox PH
## Link function: identity
##
## Formula:
## personyrs ~ age + bmi
##
## Parametric coefficients:
##     Estimate Std. Error z value Pr(>|z|)
## age 0.049473   0.003419   14.47  < 2e-16 ***
## bmi 0.018798   0.003405    5.52 3.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Deviance explained = 1.47%
## -REML =  27890  Scale est. = 1         n = 6696
```

Question 3

Describe the results from the above regression, in terms of significance.

Both terms in the model, *age* and *bmi*, are statistically significant. Each term is positively associated with the response: time since diagnosis with colon cancer.

Question 4

Run a Cox regression with *age* and *bmi* entering as smooth terms.

```
cox_mod_two <- mgcv::gam(personyrs ~ s(age) + s(bmi)
                         , weights=colon_cc
                         , data=colon
                         , family="cox.ph"
                         )
mgcv::summary.gam(cox_mod_two)
```

```
##
## Family: Cox PH
## Link function: identity
##
## Formula:
## personyrs ~ s(age) + s(bmi)
##
## Approximate significance of smooth terms:
##          edf Ref.df Chi.sq  p-value
## s(age) 2.380  2.979 209.14  < 2e-16 ***
## s(bmi) 3.022  3.707  42.86 1.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Deviance explained = 1.63%
## -REML =  27881  Scale est. = 1         n = 6696
```

Question 5

What is statistically significant?

As before, both terms appear to be statistically significant.

Question 6

Try to compare the models using your favorite method.

```
anova(cox_mod, cox_mod_two, test="Chisq")
```

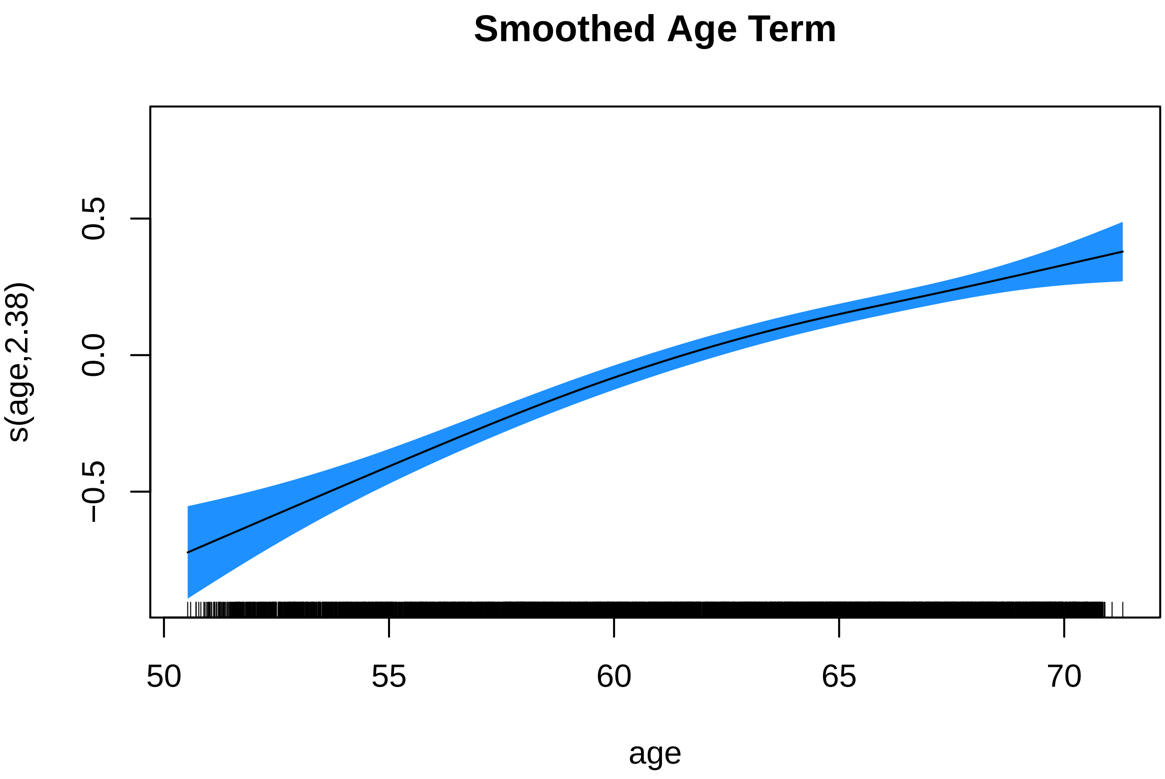
```
## Analysis of Deviance Table
##
## Model 1: personyrs ~ age + bmi
## Model 2: personyrs ~ s(age) + s(bmi)
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1      6694      55761
## 2      6688      55740 5.9715   20.681 0.002048 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The significant test statistic indicates appears that smoothing the terms is necessary.

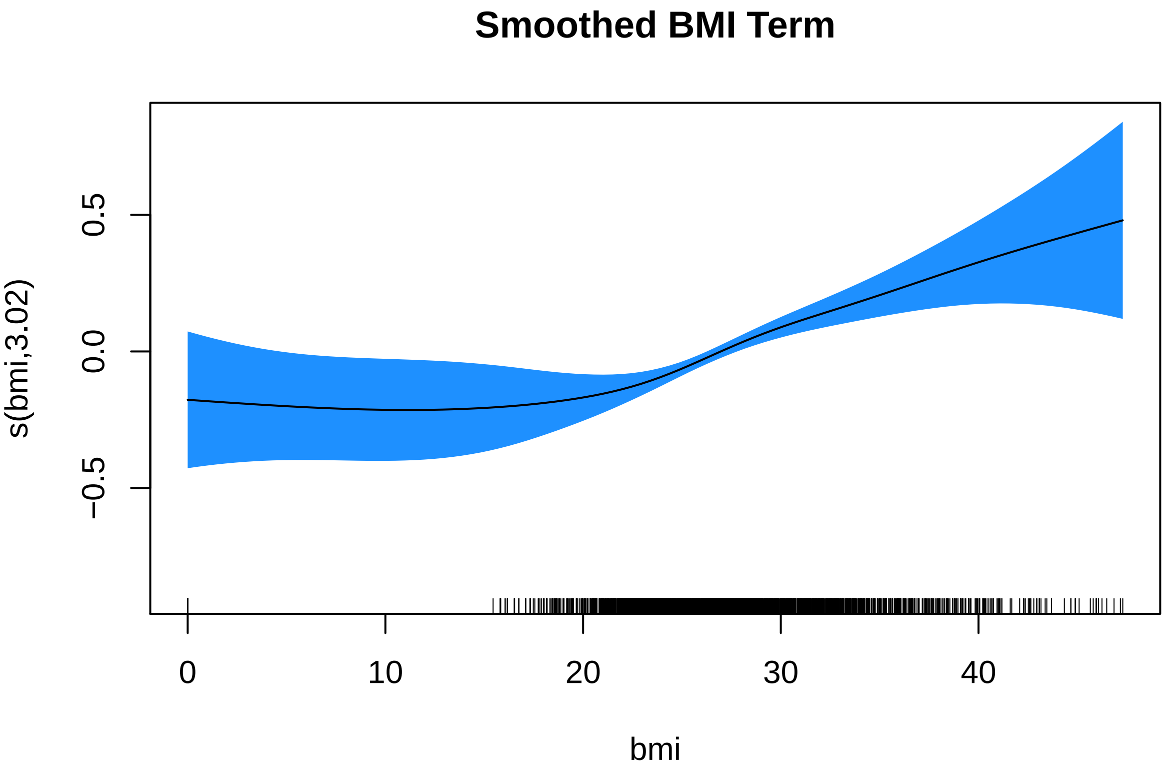
Question 7

Plot the smooths, their pointwise confidence intervals, and the Cox residuals.

```
# plot the smoothed terms first

mgcv::plot.gam(cox_mod_two, shade=T, shade.col='dodger blue', select=1
               , main="Smoothed Age Term")
```

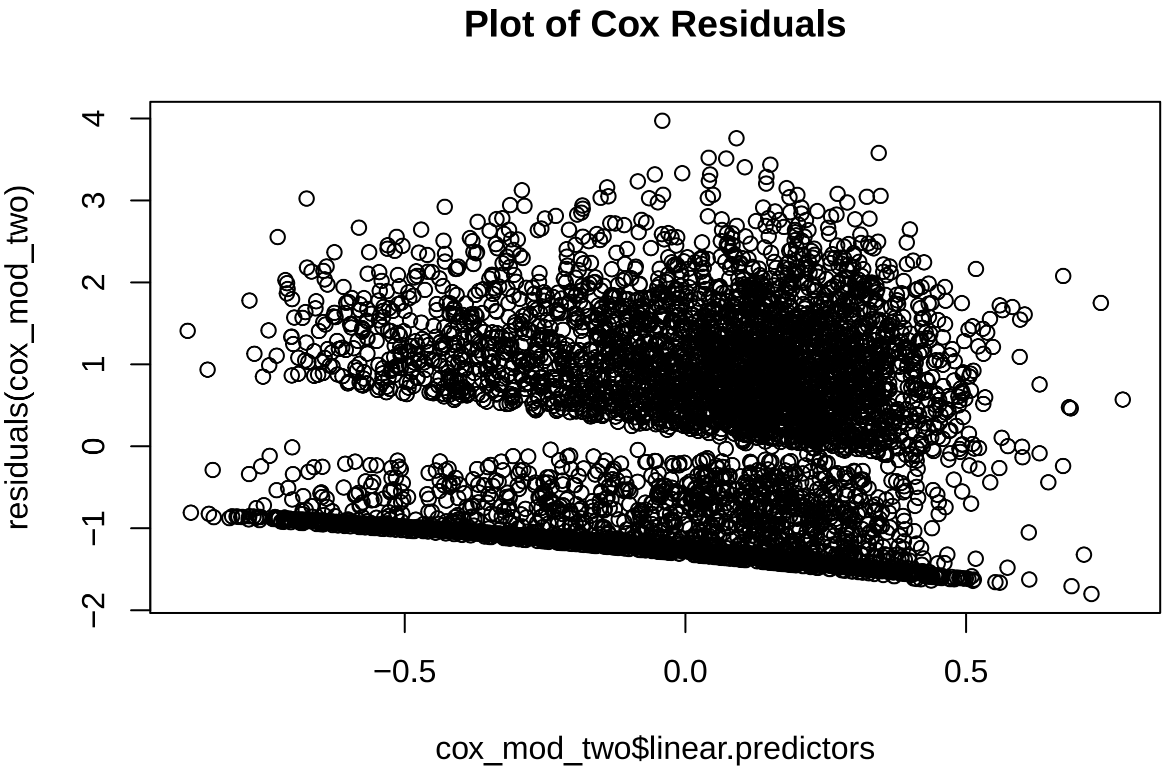


```
mgcv::plot.gam(cox_mod_two, shade=T, shade.col='dodger blue', select=2
              , main="Smoothed BMI Term")
```



```
# plot the Cox residuals
plot(cox_mod_two$linear.predictors, residuals(cox_mod_two)
     , main="Plot of Cox Residuals"
```

```
    )
```



Question 8

Do the Cox residuals look something like the Simon Wood data?

Not especially. It looks like we have distinct patterns in our residual plots, and may need to look for additional predictors.

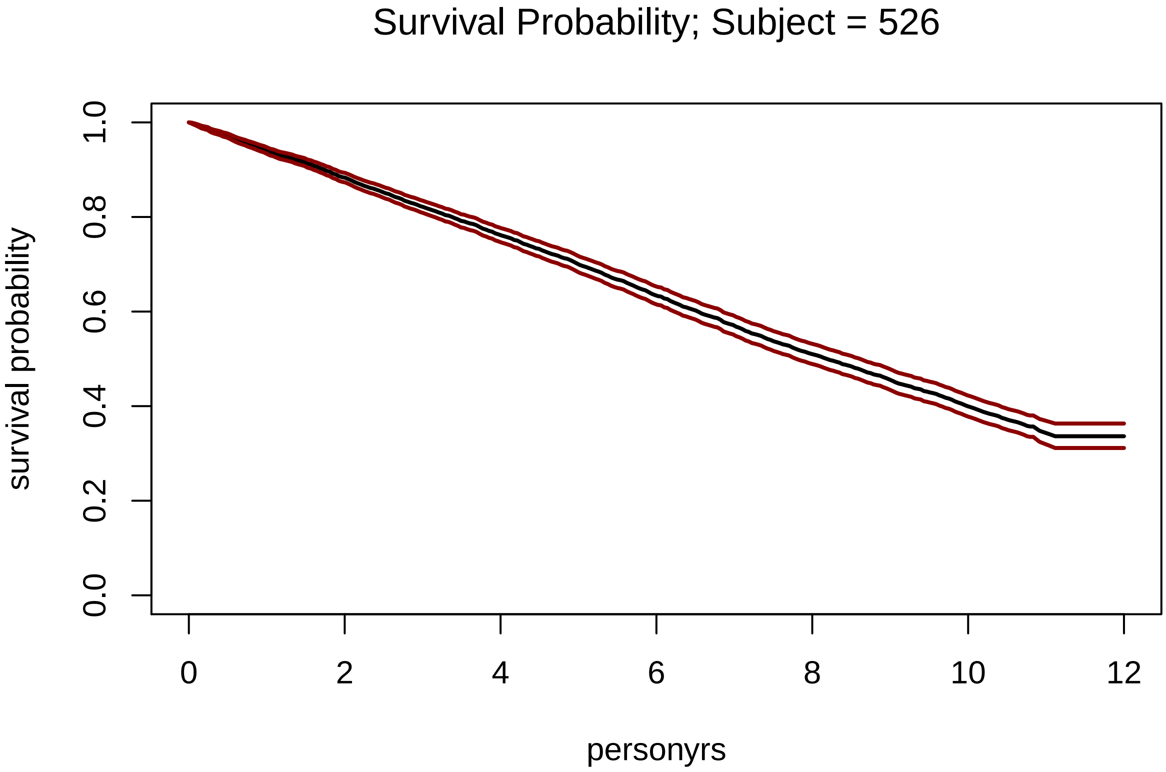
Question 9

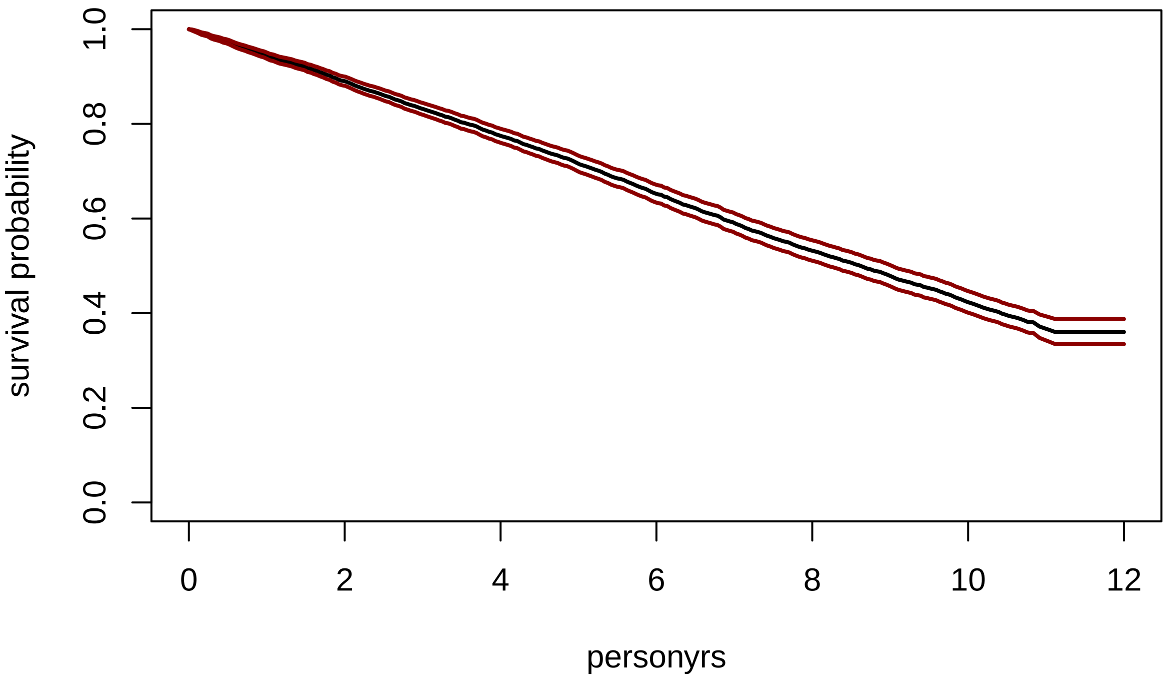
Pick any four people and plot their survivial curves and pointwise confidence intervals.

It will be easiest to write a function for this that adapts the Simon Wood code to generalize somewhat.

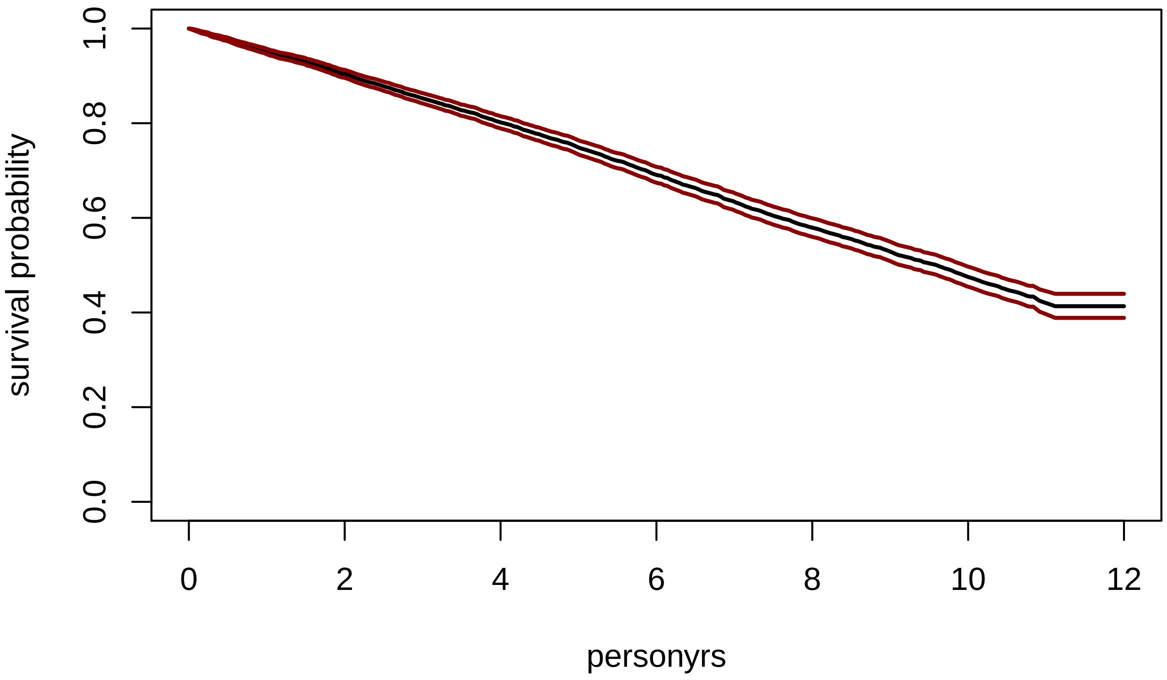
```
survival_plotter <- function(subjectnum=1, np=300, dset, mod){
                        newd <- data.frame(matrix(0, np, 0))
                        for (n in names(dset))
                          newd[[n]] <- rep(dset[[n]][subjectnum], np)

                        newd$personyrs <- seq(0, 12, length=np)

                        fv <- predict(mod, newdata=newd, type="response", se=TRUE)

                        plot(newd$personyrs, fv$fit, type="l", lwd=2
                             , xlim=c(0, 12), ylim=c(0, 1)
                             , xlab="personyrs", ylab="survival probability"
                             , main = substitute(paste("Survival Probability; Subject = ",m)
                                                 ,list(m = subjectnum))
                        )

                        se <- fv$se.fit / fv$fit
                        lines(newd$personyrs, exp(log(fv$fit)+2*se), col="red4", lwd=2)
                        lines(newd$personyrs, exp(log(fv$fit)-2*se), col="red4", lwd=2)
                        }
```

```
# pick random people to plot
set.seed(1740)
cases <- sort(ceiling(runif(4, 0, nrow(colon))))
for (i in 1:length(cases))
  survival_plotter(subjectnum=cases[i], np=300, dset=colon, mod=cox_mod_two)
```

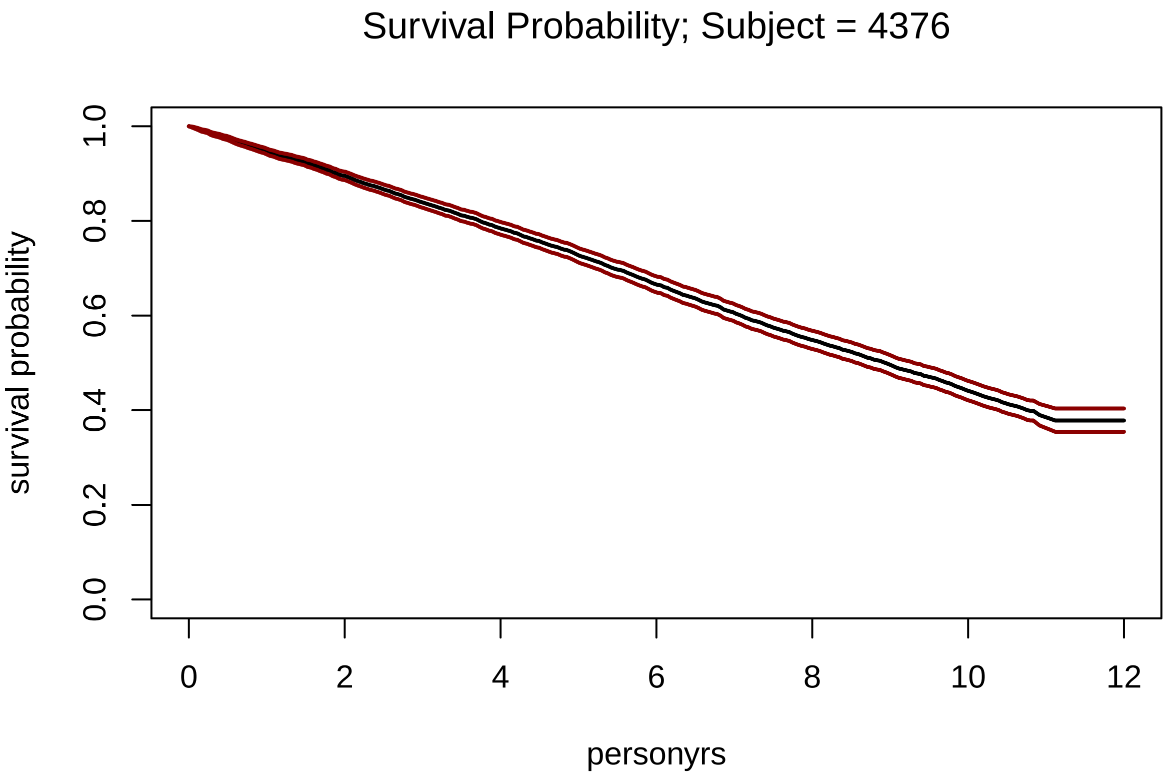




Survival Probability; Subject = 556



Survival Probability; Subject = 3618


Survival Probability; Subject = 4376
survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
personyrs