####

####This program analyzes the colon cancer data set. Model used is based out of Cox regression with splines using GAM

####colon\_cc: = 0 if the person survives colon cancer, i.e., is censored. = 1 otherwise

####nonsmoker: = 1 if the person is a nonsmoker, = 0 if they smoke

####age: Self-explanatory

####bmi: Self-explanatory

####personyrs: Time since being diagnosed with colon cancer : This is the t in hazard function h(t) being modelled as Cox

####

#### This is where my dataset is located.

setwd("F:/vigupta/OneDrive/Learning/DataScience/Statistics Texas A&M University/689/Assignment/Assignment13")

colonCancer = read.csv("Colon\_Cancer\_Data.csv")

attach(colonCancer)

#### Might not be needed but we will do anyways to be safe to repro.

set.seed(12234)

####libraries needed

library(mgcv)

####1. What percentage died ?

pctDied = 100\*sum(colonCancer$colon\_cc) / nrow(colonCancer)

##50%

####2. Run a Cox regression with age and bmi entering linearly.

##Acceptable normal fit , no separation seen between the classes

plot(density(colonCancer$age))

shapiro.test(sample(colonCancer$age , 4000))

plot(age , colon\_cc , col = colon\_cc + 1)

##Evidence of multi modality , no separation seen between the classes

plot(density(colonCancer$bmi))

plot(bmi , colon\_cc , col = colon\_cc + 1)

colFit.Linear = gam(personyrs ~ age + bmi + nonsmoker , weights = colon\_cc , family = cox.ph , data = colonCancer)

summary(colFit.Linear)

###Fit

####Family: Cox PH

####Link function: identity

####

####Formula:

####personyrs ~ age + bmi + nonsmoker

####

####Parametric coefficients:

#### Estimate Std. Error z value Pr(>|z|)

####age 0.049705 0.003418 14.541 < 2e-16 \*\*\*

####bmi 0.018054 0.003393 5.321 1.03e-07 \*\*\*

####nonsmoker -0.178275 0.034771 -5.127 2.94e-07 \*\*\*

####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####

####

####Deviance explained = 1.61%

####-REML = 27879 Scale est. = 1 n = 6696

#### We do significant effect of smoker, age , bmi,

names(colFit.Linear)

logLik.gam(colFit.Linear)

##'log Lik.' -27867.43 (df=3)

(colFit.Linear$deviance)

(colFit.Gam$deviance)

####3. Describe the results in problem 1, as a statistical significance

plot(colFit.Linear$linear.predictors,residuals(colFit.Linear),xlab="Linear predictors",

ylab = "Cox Residuals")

##' The above model we have entered age , bmi and smoking status as continuous random variables.

##' Smoking status is in (0, 1) and the coeff should be interpreted as such. 1 if non - smoker.

##'

##' Age : We see from the dataset that the age is ~ between 50 - 80 , with the coef in survival

##' fit to be + , 0.049705. Thus as we expected , with the increase in the age, risk of dying with

##' colon cancer increases. The estimated effect of age is 0.049705 with a highly significant

##' pvalue of ~ 0

##'

##' smoking : We note that the Smokers are referenced as 0 in the data set. Since the coeff is -ve , the

##' risk of a smoker having colon cancer is higher than that of non smoker

##' The estimated effect of smoking is -0.178275 with a highly significant pvalue of ~ 0

##'

##' BMI : As with age, we notice that as the BMI level increases the risk of dying also increases.

##' The estimated effect of bmi is 0.018054 with a highly significant pvalue of ~ 0

##'

##'

####4. Cox regression with age and bmi entering as smooths

colFit.Gam = gam(personyrs ~ s(age , bs = "cr" , k = 20) + s(bmi, bs = "cr" , k = 20) + nonsmoker

, weights = colon\_cc , family = cox.ph , data = colonCancer)

summary(colFit.Gam)

####Family: Cox PH

####Link function: identity

####

####Formula:

####personyrs ~ s(age, bs = "cr", k = 20) + s(bmi, bs = "cr", k = 20) +

#### nonsmoker

####

####Parametric coefficients:

#### Estimate Std. Error z value Pr(>|z|)

####nonsmoker -0.1744 0.0348 -5.011 5.42e-07 \*\*\*

####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####

####Approximate significance of smooth terms:

#### edf Ref.df Chi.sq p-value

####s(age) 2.402 3.013 210.98 < 2e-16 \*\*\*

####s(bmi) 2.887 3.554 39.27 6.62e-08 \*\*\*

####---

####Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

####

####Deviance explained = 1.76%

####-REML = 27868 Scale est. = 1 n = 6696

logLik.gam(colFit.Gam)

##'log Lik.' -27857.93 (df=7.567786)

##'

####5. What is statistically significant

##' We note that as with the linear age and bmi , the splines on age and bmi

##' are also statistically significant with a p value of ~ 0 for both the

##' covariates.

####6. Try to compare the models in Problems 4 and 1 using your favorite method

### Compare the two fits via a Chisq test on anova.

anova(colFit.Gam , colFit.Linear , test = "Chisq")

###Analysis of Deviance Table

###

###Model 1: personyrs ~ s(age, bs = "cr", k = 20) + s(bmi, bs = "cr", k = 20) +

### nonsmoker

###Model 2: personyrs ~ age + bmi + nonsmoker

### Resid. Df Resid. Dev Df Deviance Pr(>Chi)

###1 6687.2 55716

###2 6693.0 55735 -5.8469 -19.01 0.003709 \*\*

###---

###Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#### We clearly see that the spline fits are statistically significant.

####7. Plot the smooths and their pointwise CI and the Cox residuals.

#### Plotting the smooths with the fit

par(mfrow = c(2,1))

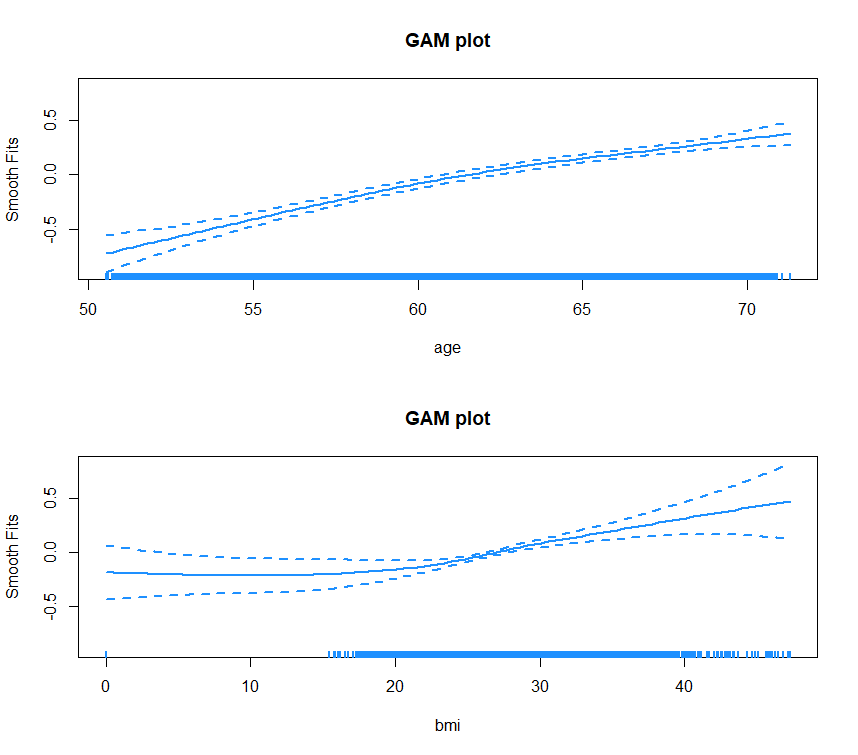
plot(colFit.Gam , col = "dodgerblue" , lwd = 2 , ylab = "Smooth Fits" , main = "GAM plot")

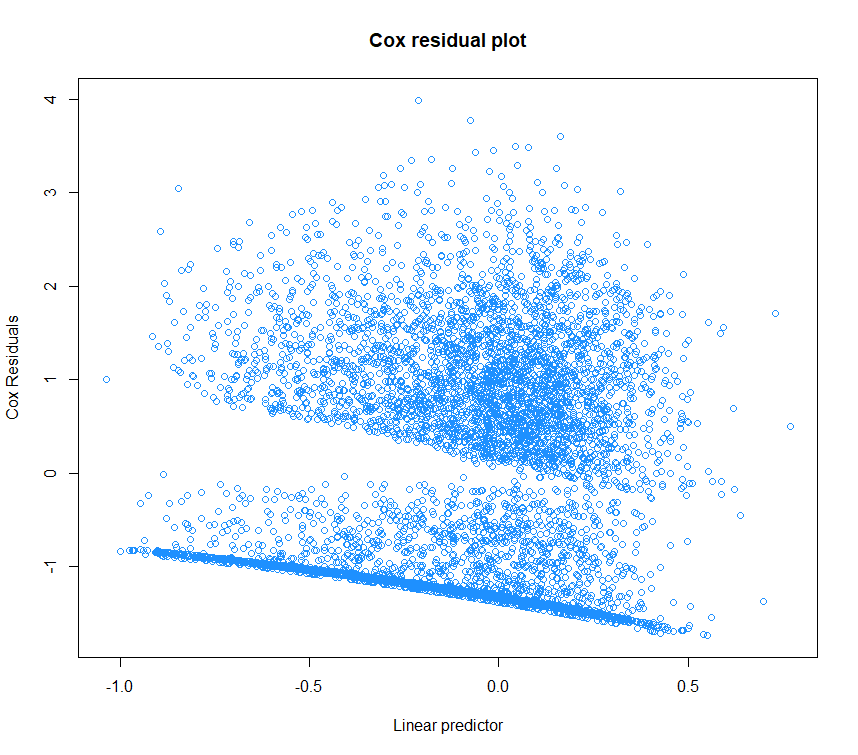
#### Plotting the cox residuals

par(mfrow = c(1,1))

plot(colFit.Gam$linear.predictors , residuals(colFit.Gam) , xlab = "Linear predictor" ,

ylab = "Cox Residuals" , main = "Cox residual plot" , col = "dodgerblue" )





####8. Do the Cox residuals look something like the Simon Wood data?

#### Yes the COX residuals between Simon Wood and COlon cancer models look similar.

####9. Pick any 4 people and plot their survival curves and pointwise confidence intervals.

set.seed(12234)

ng = 100

samples = sample(1:nrow(colonCancer) , 4)

nonsmoker[samples]

agegrid= seq(min(age) , max(age) , length = ng )

bmigrid= seq(min(bmi) , max(bmi) , length = ng )

yearsgrid= seq(min(personyrs) , max(personyrs) , length = ng )

par(mfrow = c(2,2))

for (i in samples) {

nonsmokegrid = rep(nonsmoker[i] , ng)

newdf = data.frame("nonsmoker" = nonsmokegrid, "age" = agegrid ,"bmi" = bmigrid, "personyrs" = yearsgrid )

head(newdf)

head(colonCancer)

pr = predict(colFit.Linear,newdata=newdf,type="response",se=TRUE)

plot(newdf$personyrs,pr$fit,type="l",ylim=c(0.,1),xlab="Person Years",

ylab="Risk to Die ",lwd=2,col="blue",

main = substitute(paste("Survival Curve for Subject = ",m , ", Nonsmoker = " , s),list(m = i , s = nonsmoker[i])))

legend("bottomleft" , c("Mean func" , "C.I - mean") , text.col = c("blue" , "red4"))

## Add intervals based on cumulative hazard s.e...

# This appears to be correct, but involves heavy theory

se = pr$se.fit/pr$fit

lines(newdf$personyrs,exp(log(pr$fit)+2\*se),col="red4",lwd=2 , lty = 2)

lines(newdf$personyrs,exp(log(pr$fit)-2\*se),col="red4",lwd=2 , lty = 2)

}

