####

####The Framingham Heart Study https://www.framinghamheartstudy.org/ is the longest on-going

####study of risk factors from heart disease and many other types of chronic diseases. It is a national

####treasure started in 1948.

####Framingham data contains observations for medical test on systolic blood pressure

####VARIABLE NAMES and DEFINITIONS:

####COL 01 -- OBS = observation number (1-1615)

####COL 02 -- AGE = age at exam 2

####COL 03 -- SBP21 = first systolic blood pressure at exam 2

####COL 04 -- SBP22 = second systolic blood pressure at exam 2

####COL 05 -- SBP31 = first systolic blood pressure at exam 3

####COL 06 -- SBP32 = second systolic blood pressure at exam 3

####COL 07 -- SMOKE = present smoking at exam 1

####COL 08 -- CHOLEST2 = serum cholesterol at exam 2

####COL 09 -- CHOLEST3 = serum cholesterol at exam 3

####COL 10 -- FIRSTCHD = indicator of first evidence of CHD occurring at exam 3

####through 6, i.e., within an eight-year follow-up period

####to exam 2. CHD = coronary heart disease

####GENERAL INFORMATION:

####1. The data are for MALES only.

####2. The data contain complete records only.

####This exercise is to study whether systolic blood pressure and serum cholesterol are related, and

####whether smoking status (and later on age) are factors in systolic blood pressure (it is).

####Vivek Kumar Gupta , Stat 689 Assignment 7

####

####Set the working director for the assignment

setwd("F:/OneDrive/Learning/DataScience/Statistics Texas A&M University/689/Assignment/Assignment7")

#Clear the working spaace and load in the libraries

rm(list = ls())

set.seed(1234)

library("HRW")

library("mgcv")

library("nlme")

## Read the data into a frame called framingham

framingham = read.csv("Framingham(1).csv")

####There are four systolic blood pressure measurements. Take their average and create the variable

####Average the two cholesterol measurements and take their logarithm

framingham$LSBP = log(apply(framingham[,c("SBP22" , "SBP21" ,"SBP31" ,"SBP32")], 1,mean) - 50)

framingham$Lcholest = log(apply(framingham[,c("Cholest2" , "Cholest3" )], 1,mean))

#####Reset the dataframe

framingham = framingham[, c("Age" , "Smoker" , "LSBP" , "Lcholest" , "CHD")]

####Convert Smoker column to factors and check the reference

framingham$Smoker = as.factor(framingham$Smoker)

unique(framingham$Smoker)

####Attach the frame so that it is easier to reference

attach(framingham)

framingham.fit.gam.logistic.nosplines = gam(CHD ~ Age + Smoker + Lcholest + LSBP

, family = binomial (link = "logit"))

**################Answer to Question 1 ####################**

####Fit a logistic gam with only LSBP modeled as a spline. Quote the p-values

####for all 4 predictors, and answer whether the fit suggest that LSBP should be modeled as a spline.

framingham.fit.gam.logistic.lspbspline = gam(CHD ~ Age + Smoker + Lcholest + s(LSBP , bs = "cr" , k = 23)

, family = binomial (link = "logit"))

summary(framingham.fit.gam.logistic.lspbspline)

####Family: binomial

####Link function: logit

####

####Formula:

####CHD ~ Age + Smoker + Lcholest + s(LSBP, bs = "cr", k = 23)

####

####Parametric coefficients:

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

####(Intercept) -20.33972 3.28988 -6.183 6.31e-10 \*\*\*

####Age 0.05673 0.01190 4.767 1.87e-06 \*\*\*

####Smoker1 0.60475 0.25094 2.410 0.016 \*

####Lcholest 2.67958 0.57843 4.633 3.61e-06 \*\*\*

####---

####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(LSBP) 1.748 2.22 16.07 0.000547 \*\*\*

####---

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

####

####R-sq.(adj) = 0.0433 Deviance explained = 9.17%

####UBRE = -0.48979 Scale est. = 1 n = 1615

**####We can see a significant spline coeff indicating that LSBP should be fit as a spline.**

####Question8 part

####Pvalue of the smoking variable is 0.016, which is significant. The estimate implies that

#### odds of CVD for a smoker versus a nonsmoker is exp(0.60475) 1.830794 while controlling for other variables.

####Question9 part , 95% CI of the OR estimate is

exp(0.60475 + c(-1 , 1) \* qnorm(1 - .05 / 2) \* 0.25094)

1.119539 2.993919

**################Answer to Question 2 ####################**

####Fit a logistic gam with only Lcholest modeled as a spline.

####Quote the p-values for all 4 predictors, and answer whether the fit suggest that Lcholest should be modeled as a spline.

framingham.fit.gam.logistic.Lcholestspline = gam(CHD ~ Age + Smoker + LSBP + s(Lcholest , bs = "cr" , k = 23)

, family = binomial (link = "logit"))

summary(framingham.fit.gam.logistic.Lcholestspline)

####Formula:

####CHD ~ Age + Smoker + LSBP + s(Lcholest, bs = "cr", k = 23)

####

####Parametric coefficients:

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

####(Intercept) -13.04593 1.85195 -7.044 1.86e-12 \*\*\*

####Age 0.05580 0.01188 4.697 2.64e-06 \*\*\*

####Smoker1 0.60900 0.25134 2.423 0.0154 \*

####LSBP 1.66355 0.42118 3.950 7.82e-05 \*\*\*

####---

####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(Lcholest) 1.003 1.006 22.48 2.23e-06 \*\*\*

####---

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

####

####R-sq.(adj) = 0.042 Deviance explained = 8.92%

####UBRE = -0.48931 Scale est. = 1 n = 1615

**####We can see a significant spline coeff indicating that LCholest should be fit as a spline.**

####Question8 part

####Pvalue of the smoking variable is 0.60900, which is significant. The estimate implies that

#### odds of CVD for a smoker versus a nonsmoker is exp(0.60900) 1.838592 while controlling for other variables.

####Question9 part , 95% CI of the OR estimate is

exp(0.60900 + c(-1 , 1) \* qnorm(1 - .05 / 2) \* 0.25134)

1.123426 3.009029

**################Answer to Question 3 ####################**

####Fit a logistic gam with only age modeled as a spline. Quote the p-values for all 4 predictors,

####and answer whether the fit suggest that age should be modeled as a spline.

framingham.fit.gam.logistic.Agespline = gam(CHD ~ Lcholest + Smoker + LSBP + s(Age , bs = "cr" , k = 23)

, family = binomial (link = "logit"))

summary(framingham.fit.gam.logistic.Agespline)

####Formula:

####CHD ~ Lcholest + Smoker + LSBP + s(Age, bs = "cr", k = 23)

####

####Parametric coefficients:

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

####(Intercept) -25.1600 3.6525 -6.888 5.64e-12 \*\*\*

####Lcholest 2.6873 0.5815 4.622 3.81e-06 \*\*\*

####Smoker1 0.5955 0.2510 2.372 0.0177 \*

####LSBP 1.6861 0.4214 4.001 6.30e-05 \*\*\*

####---

####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) 2.255 2.83 22.88 4.63e-05 \*\*\*

####---

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

**####We can see a significant spline coeff indicating that Age should be fit as a spline.**

####Question8 part

####Pvalue of the smoking variable is 0.5955, which is significant. The estimate implies that

#### odds of CVD for a smoker versus a nonsmoker is exp(0.5955) 1.813938 while controlling for other variables.

####Question9 part , 95% CI of the OR estimate is

exp(0.5955 + c(-1 , 1) \* qnorm(1 - .05 / 2) \* 0.2510)

1.109100 2.966702

**################Answer to Question 4 ####################**

####Fit a logistic gam with LSBP, Lcholest and age modeled as splines.

####Quote the p-values for all 4 predictors, and tell me which of the spline terms seem

####like they are worth modeling as a spline

framingham.fit.gam.logistic.spline = gam(CHD ~ s(Lcholest , bs = "cr" , k = 23) + Smoker + s(LSBP , bs = "cr" , k = 23) + s(Age , bs = "cr" , k = 23)

, family = binomial (link = "logit"))

summary(framingham.fit.gam.logistic.spline)

####Family: binomial

####Link function: logit

####

####Formula:

####CHD ~ s(Lcholest, bs = "cr", k = 23) + Smoker + s(LSBP, bs = "cr",

#### k = 23) + s(Age, bs = "cr", k = 23)

####

####Parametric coefficients:

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

####(Intercept) -3.2512 0.2457 -13.230 <2e-16 \*\*\*

####Smoker1 0.5925 0.2507 2.363 0.0181 \*

####---

####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(Lcholest) 1.002 1.004 20.62 5.74e-06 \*\*\*

####s(LSBP) 1.638 2.065 16.23 0.000361 \*\*\*

####s(Age) 2.194 2.754 23.32 3.70e-05 \*\*\*

####---

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

####

####R-sq.(adj) = 0.0446 Deviance explained = 9.71%

####UBRE = -0.4914 Scale est. = 1 n = 1615

**####We see that all the three coeffs are significant hence the fit suggests that we should model all three covariates**

**####LCholest LSBP and Age as splines.**

####Question8 part

####Pvalue of the smoking variable is 0.5925, which is significant. The estimate implies that

#### odds of CVD for a smoker versus a nonsmoker is exp(0.5925) 1.808504 while controlling for other variables.

####Question9 part , 95% CI of the OR estimate is

exp(0.5925 + c(-1 , 1) \* qnorm(1 - .05 / 2) \* 0.2507)

1.106428 2.956077

**################Answer to Question 5 ####################**

####Since all the three covariates have significant spline coeffs , we will use all three in the model

#### hence our model framingham.fit.gam.logistic.spline stays

**################Answer to Question 6 ####################**

anova( framingham.fit.gam.logistic.nosplines , framingham.fit.gam.logistic.spline, test = "Chisq" )

####Analysis of Deviance Table

####

####Model 1: CHD ~ Age + Smoker + Lcholest + LSBP

####Model 2: CHD ~ s(Lcholest, bs = "cr", k = 23) + Smoker + s(LSBP, bs = "cr",

#### k = 23) + s(Age, bs = "cr", k = 23)

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

####1 1610.0 814.76

####2 1607.2 807.72 2.8228 7.0493 0.06152 .

**#### The p value for anova comparision is 0.06152 , which indicates that the spline fit may not be very significant**

**#### than the simple logistic fit with non of the covariates modelled as splines.**

**################Answer to Question 7 ####################**

dev.off()

par(mfrow = (c(2,2)))

####plot(framingham.fit.gam.logistic.Lcholestspline,shade = TRUE,shade.col = "palegreen",

#### select = 1,

#### ylab = "Logit of Prob of CVD",

#### xlab = "Lcholest",

#### main = "Fit for Lcholest - Link Scale",

#### rug = FALSE)

####rug(framingham$Lcholest,col = "dodgerblue",quiet = TRUE)

####

####plot(framingham.fit.gam.logistic.Lcholestspline,shade = TRUE,shade.col = "palegreen",

#### trans = plogis,scale = FALSE,select = 1,

#### ylab = "Probability of CVD",

#### xlab = "Lcholest"

#### ,main = "Fit for Lcholest - response scale",rug = FALSE)

####rug(framingham$Lcholest,col = "dodgerblue",quiet = TRUE)

plot(framingham.fit.gam.logistic.spline,shade = TRUE,shade.col = "palegreen",

select = 1,

ylab = "Logit of Prob of CVD",

#xlab = "LSBP",

main = "Fit for LCholest - Link Scale",

rug = FALSE)

rug(framingham$Lcholest,col = "dodgerblue",quiet = TRUE)

plot(framingham.fit.gam.logistic.spline,shade = TRUE,shade.col = "palegreen",

trans = plogis,scale = FALSE,select = 1,

ylab = "Probability of CVD",

#xlab = "LCholest"

main = "Fit for LCholest - response scale",rug = FALSE)

rug(framingham$Lcholest,col = "dodgerblue",quiet = TRUE)

plot(framingham.fit.gam.logistic.spline,shade = TRUE,shade.col = "palegreen",

select = 2,

ylab = "Logit of Prob of CVD",

#xlab = "LSBP",

main = "Fit for LSBP - Link Scale",

rug = FALSE)

rug(framingham$LSBP,col = "dodgerblue",quiet = TRUE)

plot(framingham.fit.gam.logistic.spline,shade = TRUE,shade.col = "palegreen",

trans = plogis,scale = FALSE,select = 2,

ylab = "Probability of CVD",

#xlab = "LCholest"

main = "Fit for LSBP - response scale",rug = FALSE)

rug(framingham$LSBP,col = "dodgerblue",quiet = TRUE)

plot(framingham.fit.gam.logistic.spline,shade = TRUE,shade.col = "palegreen",

select = 3,

ylab = "Logit of Prob of CVD",

#xlab = "LSBP",

main = "Fit for Age - Link Scale",

rug = FALSE)

rug(framingham$Age,col = "dodgerblue",quiet = TRUE)

plot(framingham.fit.gam.logistic.spline,shade = TRUE,shade.col = "palegreen",

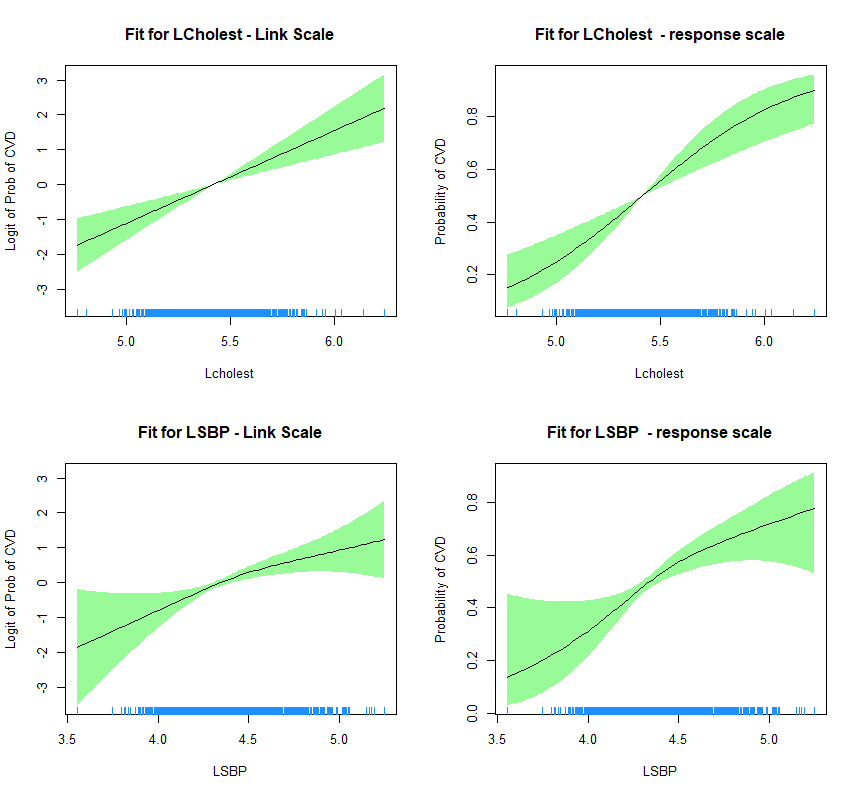
trans = plogis,scale = FALSE,select = 3,

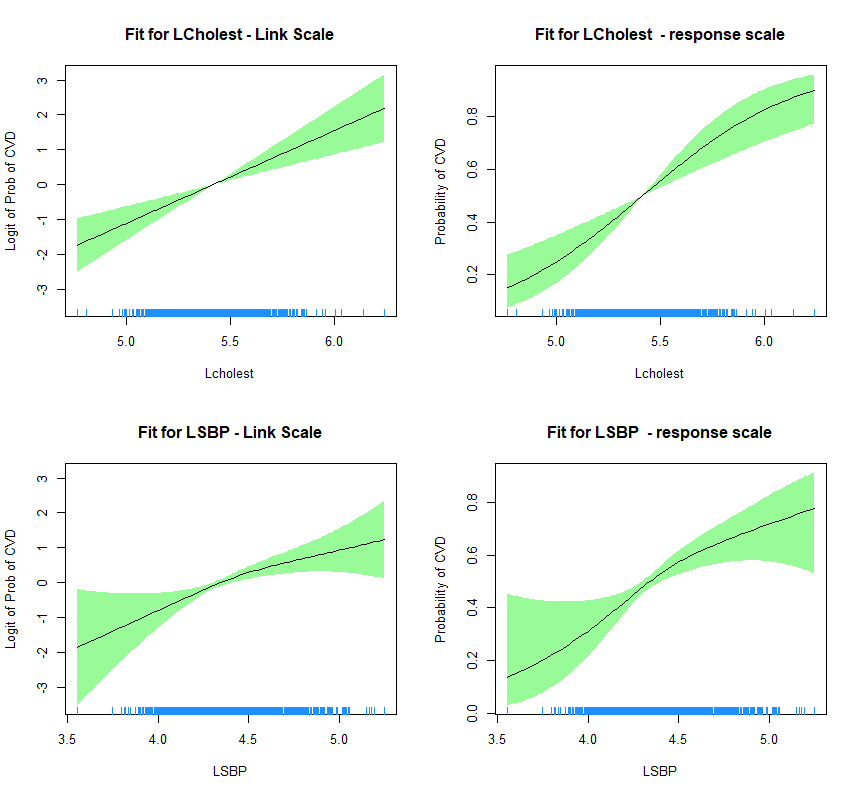
ylab = "Probability of CVD",

#xlab = "LCholest"

main = "Fit for Age - response scale",rug = FALSE)

rug(framingham$Age,col = "dodgerblue",quiet = TRUE)





**################Answer to Question 8 and 9 ####################**

#### This has been answered inline in answers to parts 1 through 4

**################Answer to Question 10####################**

#### Selecting LCholest to get the OR at min LCholest and OR at LCholest and then finding its ratio

Lcholestorder = order(Lcholest)

ratio = exp((fitted(framingham.fit.gam.logistic.spline)[Lcholestorder])[length(Lcholestorder)]) / exp((fitted(framingham.fit.gam.logistic.spline)[Lcholestorder])[1])

1.470458

**################Answer to Question 11 ####################**

#### From the analysis, it is shown that all the factors, Log of cholestrol level , log of sys blood pressure , smoking status and Age

#### are crucial in determining the risk of getting coronoary heart disease.

#### While controlling for all the factors except smoking , data reveals that odds of having a CHD is about 1.8 times higher for smoker than that of non

#### smoker.

#### It is seen that after age if 45 , the risk is of getting a CHD is higher than 0.5 and similar is observeration when Lcholest level is

#### greater than 5.5 and LSBP is greater than 4.5. These are the ranges of Lcholest , LSBP and Age at which it is more likely to get

#### a CHD.

####