#######

#######This program performs weighted analysis on the Pamlico Sound Salinity Data. We will be using

####### curated data where the leverage points are removed and will be fitting with defaults in GAM.

#######

# Clear the workspace

rm(list = ls())

set.seed(1234)

dev.off()

library("mgcv")

library("HRW")

setwd("F:/OneDrive/Learning/DataScience/Statistics Texas A&M University/689/<>

#######

####### PART 1 : Pamlico Sound Salinity Data

#######

salinity.data = read.csv("Salinity\_Data\_Modified\_Leverage\_Points.csv")

attach(salinity.data)

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

####Fit a GAM where salinity is regressed on a smooth function of discharge. Plot discharge

####(x-axis) against the absolute residuals from this fit (y-axis).

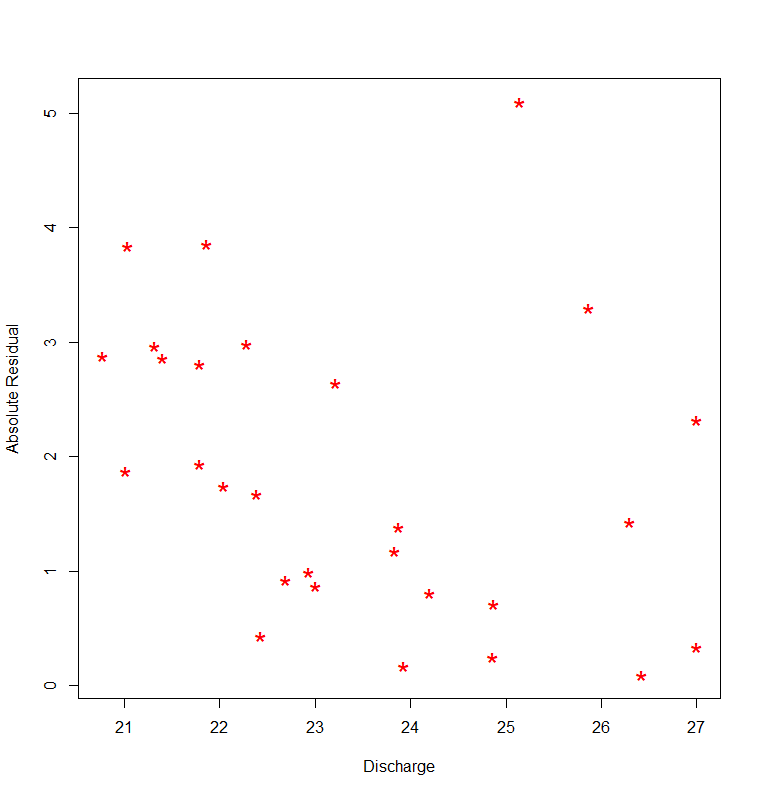
####Get the base GAM fit by regressing Salinity on the spline of discharge

salinity.gam.fit.default = gam(salinity ~ s(discharge , bs = "cr") , data = salinity.data)

ord = order(discharge)

plot(discharge[ord] , abs(salinity.gam.fit.default$residuals), cex = 2 , pch = "\*" , col = 2

, xlab = "Discharge" , ylab = "Absolute Residual")



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

####Calculate the ratio of the maximum predicted absolute residual to the minimum predicted

####absolute residual.

absresids = abs(salinity.gam.fit.default$residuals)

salinity.fitted = fitted(salinity.gam.fit.default)

gam.abs.resid.fit = gam(absresids ~ s(discharge , bs = "cr"))

ratio = (max(abs(fitted(gam.abs.resid.fit))) / min(abs(fitted(gam.abs.resid.fit))))

#5.014241

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

####Is there any concern from the calculation in (2) that heteroscedasticity might be an issue

####for pointwise confidence intervals?

####Since the ratio is greater than 3 (5.014241), there is a concern that heteroscedasticity might be an issue

####for pointwise confidence intervals

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

weight = 1 / (fitted(gam.abs.resid.fit) ^ 2)

salinity.gam.fit.default.weighted = gam(salinity ~ s(discharge , bs = "cr") , data = salinity.data , weight = weight)

plot(discharge[ord],fitted(salinity.gam.fit.default.weighted)[ord], xlab ="Dischagre" , ylab = "Fitted values " ,

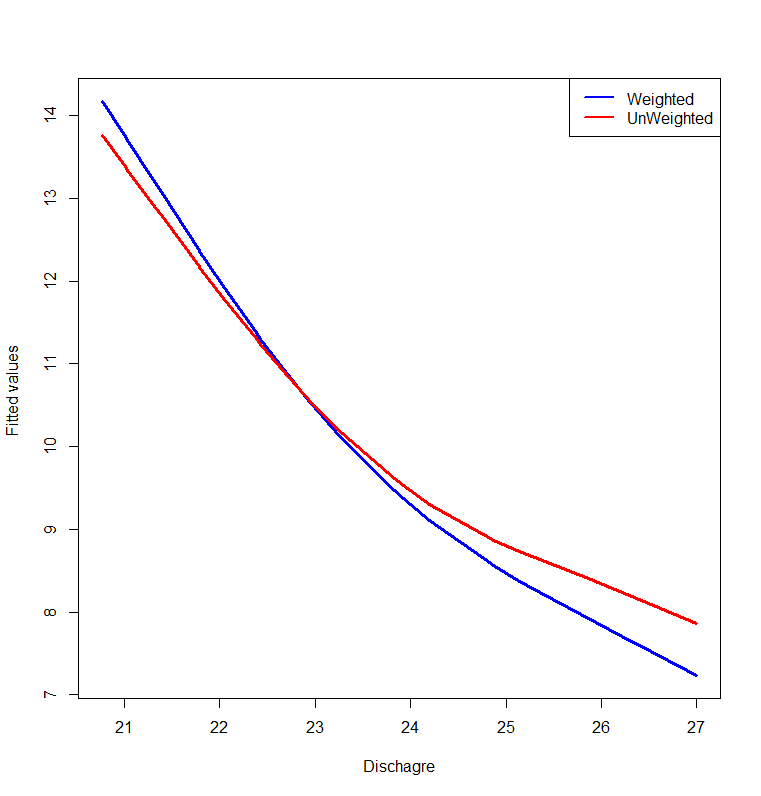
lwd=3,col="blue",type='l')

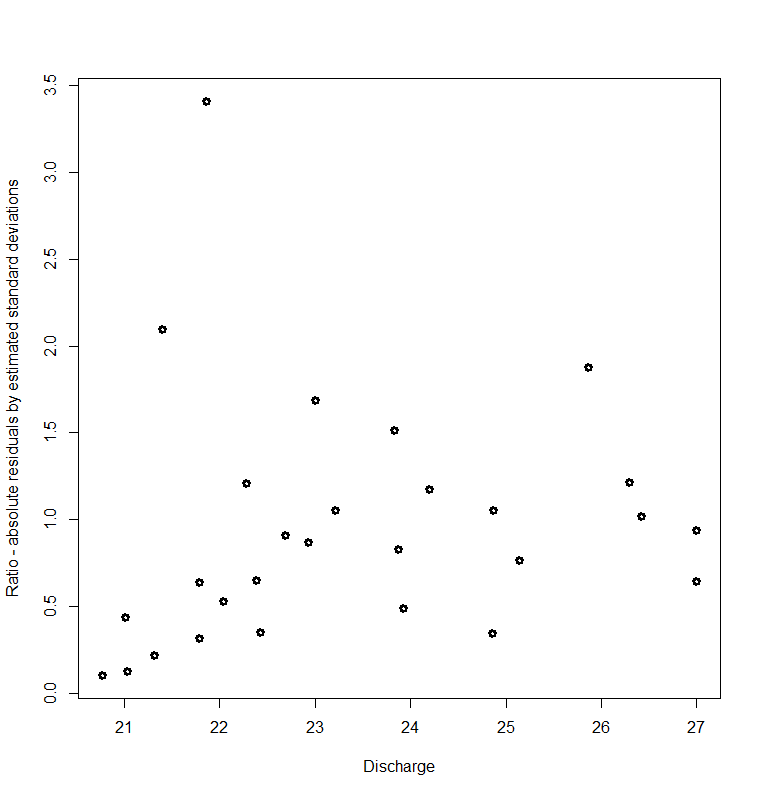
lines(discharge[ord],fitted(salinity.gam.fit.default)[ord],lwd=3,col="red",type='l')

legend("topright" , c("Weighted" , "UnWeighted") , lwd = 2 , col = c("blue" , "red"))

plot(discharge[ord],abs(salinity.gam.fit.default.weighted$residuals[ord])/fitted(gam.abs.resid.fit)[ord],lwd=3

, xlab = "Discharge" , ylab = "Ratio - absolute residuals by estimated standard deviations")





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

dev.off()

par(mfrow = c(1 ,2))

ng = 1001

dischargeGrid = seq(min(discharge) , max(discharge) , length = ng)

predDischarge.weighted = predict(salinity.gam.fit.default.weighted , newdata = data.frame(discharge = dischargeGrid)

, se = TRUE)

predDischarge.weighted.upper = predDischarge.weighted$fit + 1.96 \* predDischarge.weighted$se.fit

predDischarge.weighted.lower = predDischarge.weighted$fit - 1.96 \* predDischarge.weighted$se.fit

plot(dischargeGrid , predDischarge.weighted$fit , xlab = "Discharge" , ylab = "Predicted" , type = "l" , col = 2 , lwd = 2

, main = "Weighted")

points(dischargeGrid , predDischarge.weighted.upper , lty = 2 , col = 2 , type = "l")

points(dischargeGrid , predDischarge.weighted.lower , lty = 2 , col = 2 , type = "l")

predDischarge.unweighted = predict(salinity.gam.fit.default , newdata = data.frame(discharge = dischargeGrid)

, se = TRUE)

predDischarge.unweighted.upper = predDischarge.unweighted$fit + 1.96 \* predDischarge.unweighted$se.fit

predDischarge.unweighted.lower = predDischarge.unweighted$fit - 1.96 \* predDischarge.unweighted$se.fit

plot(dischargeGrid , predDischarge.unweighted$fit , xlab = "Discharge" , ylab = "Predicted" , type = "l" , col = 3 , lwd = 2

, main = "Unweighted")

points(dischargeGrid , predDischarge.unweighted.upper , lty = 2 , col = 3 , type = "l")

points(dischargeGrid , predDischarge.unweighted.lower , lty = 2 , col = 3 , type = "l")

#plot(salinity.gam.fit.default.weighted,shade = TRUE,shade.col = "palegreen",

# ylab = "logratiofit",

# trans = plogis,

# xlab = "Discharge",

# main = "Weighted",rug = FALSE,

# xlim=c(min(discharge),max(discharge)))

#plot(salinity.gam.fit.default,shade = TRUE,shade.col = "palegreen",

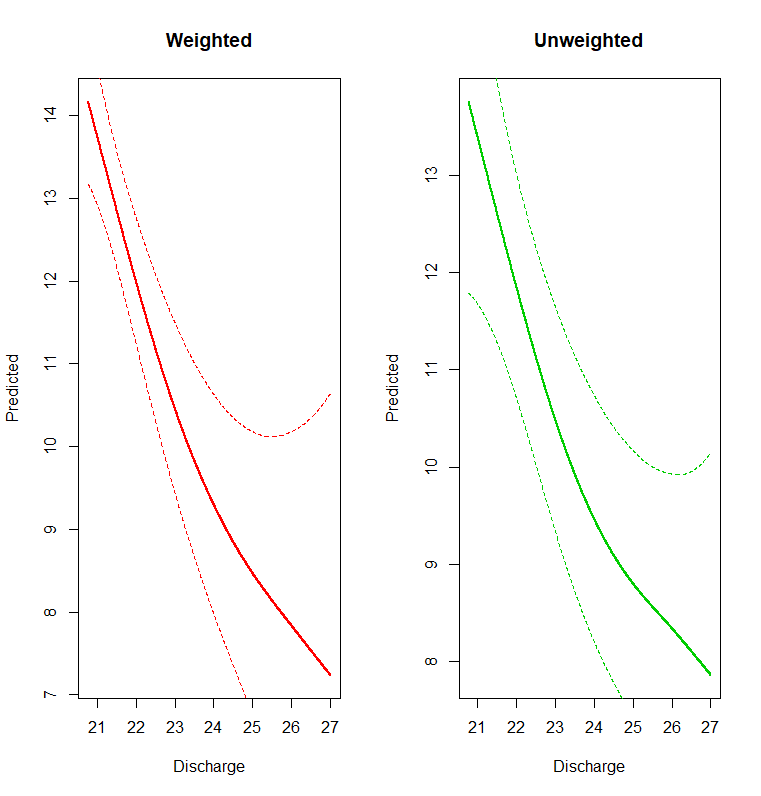
# ylab = "logratio fit",

# trans = plogis,

# xlab = "Discharge",

# main = "Unweighted",rug = FALSE,

# xlim=c(min(discharge),max(discharge)))



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

#### As we can see in the plot that for the weighted model , the CI widths for pointiwse means are narrower

#### than for the unweighted model.

#######

####### PART 2 : Blood Pressure and CHD data

#######

framingham = read.csv("Framingham\_with\_LSBP\_and\_Lcholest.csv")

attach(framingham)

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

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

dev.off()

par(mfrow = c(1, 2))

framingham.fit.factorbycurve.Int = gam(CHD ~ s(Lcholest , bs = "cr")

+ s(LSBP , bs = "cr")

+ s(Age , by = factor(Smoker) , bs = "cr")

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

framingham.fit.factorbycurve.NInt = gam(CHD ~ s(Lcholest , bs = "cr")

+ s(LSBP , bs = "cr")

+ s(Age , bs = "cr")

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

####Grid values

ng = 1001

LSBPAveg = rep(mean(LSBP) , ng)

LCholestAveg = rep(mean(Lcholest) , ng)

Agegrid = seq(min(Age) , max(Age) , length = ng)

Nsmokerg = as.factor(rep(0 , ng))

smokerg = as.factor(rep(1 , ng))

fpredNsmoker = predict(framingham.fit.factorbycurve.Int ,type = "response",

newdata = data.frame(Lcholest = LCholestAveg , LSBP = LSBPAveg , Age = Agegrid , Smoker = Nsmokerg)

, se = TRUE)

fpredNsmokerUppper = fpredNsmoker$fit + 1.96 \* fpredNsmoker$se.fit

fpredNsmokerLower = fpredNsmoker$fit - 1.96 \* fpredNsmoker$se.fit

plot(Agegrid , fpredNsmoker$fit , col = 2 , xlab = "Age" , ylab = "Probablity of CHD" , main = "Non Smoker Group" , type = "l" , lwd = 2 ,

xlim = c(min(Agegrid), max(Agegrid)))

lines(Agegrid ,fpredNsmokerUppper , col = 2 , lty = 2 )

lines(Agegrid ,fpredNsmokerLower , col = 2 , lty = 2 )

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

fpredsmoker = predict(framingham.fit.factorbycurve.Int ,type = "response",

newdata = data.frame(Lcholest = LCholestAveg , LSBP = LSBPAveg , Age = Agegrid , Smoker = smokerg)

, se = TRUE)

fpredsmokerUppper = fpredsmoker$fit + 1.96 \* fpredsmoker$se.fit

fpredsmokerLower = fpredsmoker$fit - 1.96 \* fpredsmoker$se.fit

plot(Agegrid , fpredsmoker$fit , col = 3 , xlab = "Age" , ylab = "Probablity of CHD" , main = "Smoker Group" , type = "l" , lwd = 2 ,

xlim = c(min(Agegrid), max(Agegrid)))

lines(Agegrid ,fpredsmokerUppper , col = 3 , lty = 2 )

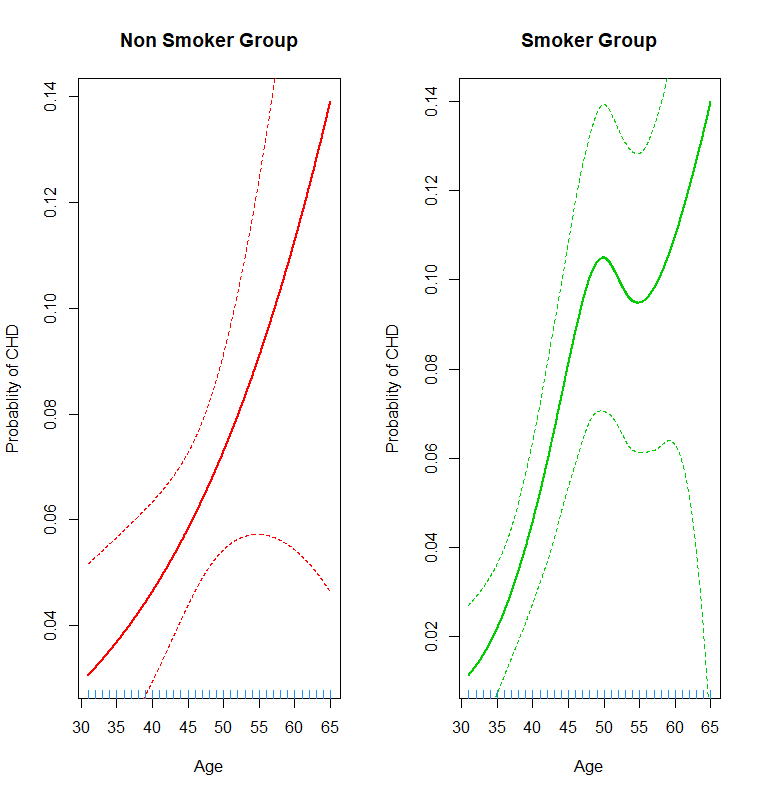
lines(Agegrid ,fpredsmokerLower , col = 3 , lty = 2 )

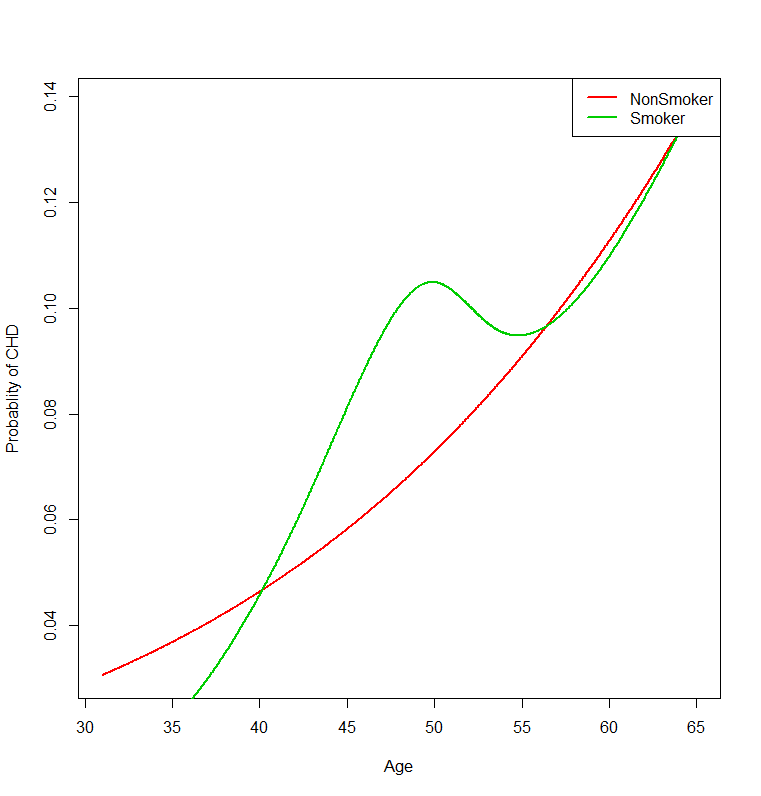
rug(Age , col = "dodgerblue",quiet = TRUE)

plot(Agegrid , fpredNsmoker$fit , col = 2 , type = "l" , lwd = 2 , xlab = "Age" , ylab = "Probablity of CHD")

points(Agegrid , fpredsmoker$fit , col = 3 , type = "l" , lwd = 2)

legend("topright" , c("NonSmoker" , "Smoker") , lwd = 2 , col = 2:3)





#######################Answer to Question 8 ####################################

anova( framingham.fit.factorbycurve.NInt, framingham.fit.factorbycurve.Int , test = "Chisq")

####Analysis of Deviance Table

####

####Model 1: CHD ~ s(Lcholest, bs = "cr") + s(LSBP, bs = "cr") + s(Age, bs = "cr")

####Model 2: CHD ~ s(Lcholest, bs = "cr") + s(LSBP, bs = "cr") + s(Age, by = factor(Smoker),

#### bs = "cr")

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

####1 1608.0 813.53

####2 1606.1 810.27 1.9504 3.2578 0.1889

####From the above anova test, we conclude that we dont have statistically significant difference in the

####model of spline of Age as factor of smoke. Low pvalue of 0.1889 does not provide any evidence in support of the theory.