###

####Fossil Data based on Geological Society of America Bulletin by Bralwer T.J et. al.

####We will be doing a simple and polynomial regression of this dataset

####The explanatory variable, X, in this dataset is Age

####The response variable is, Y, is the Strontium Ratio.

####Vivek Kumar Gupta , Stat 689 Assignment 05

####

##We first set the working directory of the data set. It is this directory where the required files are place.

setwd("F:/vigupta/OneDrive/<>")

##Clean the workspace .

rm(list = ls())

set.seed(1234)

## Read the data into a frame called fossil

fossil = read.csv("fossil(1).csv")[1:106 ,]

##Attach the dataset so that the column names can directly be used in modelling or plotting. This may be dangerous

attach (fossil)

Age = age[order(age)]

Sr.Ratio = strontium.ratio[order(age)]

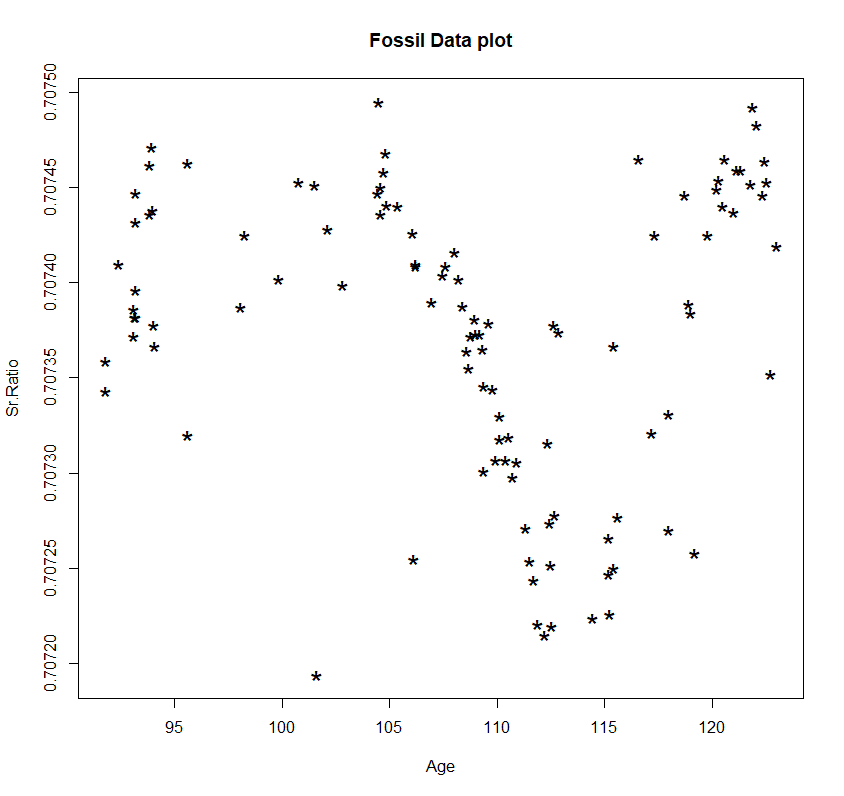
## Do some plotting to visualize the relationship between Age and Sr Ratio.

plot(Age , Sr.Ratio , pch = "\*" , cex = 2 , col = 1 , main = "Fossil Data plot")

###Interesting aspects of the relationship of the daya/

##1. The relationship between Y and X is not linear.

##2. For Age < 104 and Age > 112 the variance of the sr.ratio is pretty large.



##Fit a GAM and do a gam.check from mgcv package

library("mgcv")

fit.gam.default = gam(Sr.Ratio ~ s(Age , bs = "cr") , data = fossil)

fit.gam.k4 = gam(Sr.Ratio ~ s(Age , bs = "cr" , k = 4) , data = fossil)

fit.gam.k8 = gam(Sr.Ratio ~ s(Age , bs = "cr" , k = 8) , data = fossil )

fit.gam.k23 = gam(Sr.Ratio ~ s(Age , bs = "cr" , k = 23) , data = fossil)

## Report the p values of the fits and specify if they are significant.

#K = 4

summary(fit.gam.k4)

#pvalue of the coefficient = <2e-16 , which is statistically significant.

#K = 8

summary(fit.gam.k8)

#pvalue of the coefficient = <2e-16 , which is statistically significant.

#K = 23

summary(fit.gam.k23)

#pvalue of the coefficient = <2e-16 , which is statistically significant.

## Plot all the fits in ONE graph

plot(Age , Sr.Ratio , pch = "\*" , cex = 2 , col = 1 , main = "Fossil Data plot")

points(Age , fit.gam.k4$fitted.values , cex = 2, col = 4 , type = "l" , lwd = 2)

points(Age , fit.gam.k8$fitted.values , cex = 2, col = 5 , type = "l" , lwd = 2)

points(Age , fit.gam.k23$fitted.values , cex = 2, col = 6 , type = "l" , lwd = 2)

legend("bottomleft",

legend = c("GAM K = 4" ,"GAM K = 8","GAM K = 23"),

col = 4:6,

cex = 1.1,

pch = "\*"

)

#The fit with K = 4 as shown in the dark blue line does captures the movement of the data albeit vaguely

# i.e. to an extent . Between Age 107 to 112 it does not well capture the linear fit . Data at other points

# are highly variable in the orignial sample itself so much cannot be said on the fit to

# other cuts of the predictor.

# The fit with K = 8 shown in light blue line captures the linear fit when 107 < Age < 112 pretty close

# to as done y smooth splines

# The fit with K = 23 is a lot wiggly and is very similar to smooth spline fit.

##EDF for each of the GAM fits are shown below

#K = 4

gam.check(fit.gam.k4)

summary(fit.gam.k4)

# edf = 2.99

#K = 8

gam.check(fit.gam.k8)

summary(fit.gam.k8)

#edf = 5.98

#K = 23

gam.check(fit.gam.k23)

summary(fit.gam.k23)

#edf 15.7

##Lamdba and the p value for for each choice of K in GAM fits are shown below

#K = 4

fit.gam.k4$sp

#lam = 0.1004349

gam.check(fit.gam.k4)

#pvalue for choice of K = 4 is 0.085

#K = 8

fit.gam.k8$sp

#lam = 2.481632

gam.check(fit.gam.k8)

#pvalue for choice of K = 8 is 0.34

#K = 23

fit.gam.k23$sp

#lam = 2.001724

gam.check(fit.gam.k23)

#pvalue for choice of K = 23 is 0.99

####Some comments on the quality of the fit.

# We see from the above answers that the pvalue of K = 4 is less than 0.1

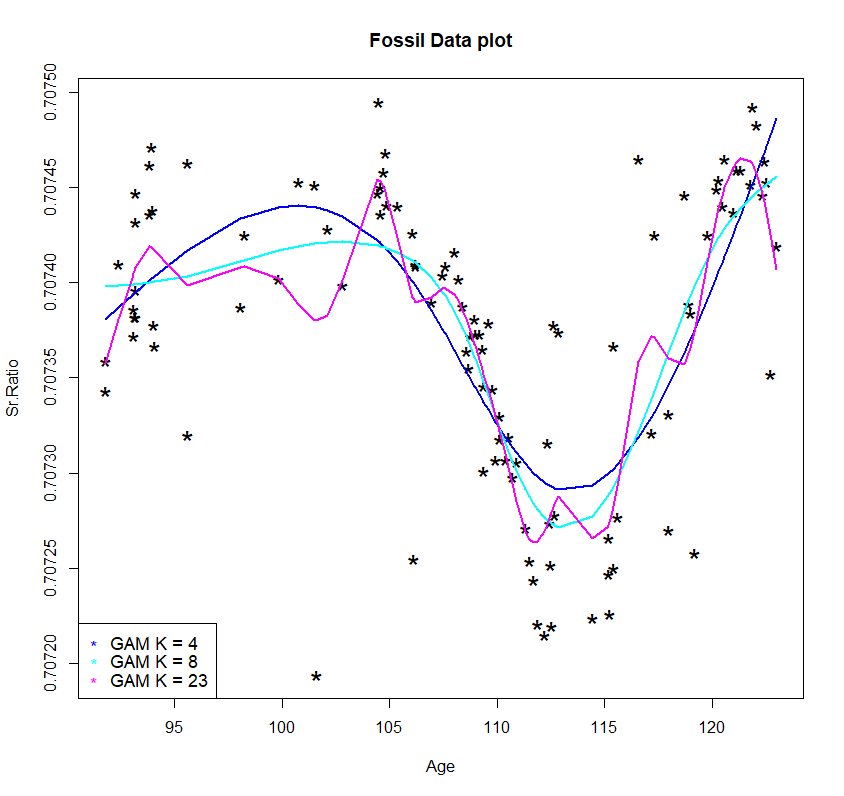
#The fit with K = 4 as shown in the dark blue line in the picture above does captures the movement of the data albeit vaguely

# i.e. to an extent . Between Age 107 to 112 it does not well capture the linear fit .Also,

# we can see that 102 < Age < 106 , gam fit with K = 4 does not captures a small hump in the data plot.

#Similar is the observation towards the higher values of Age.

# So, it does appear intuitively from the graph that K = 4 bs might bot be sufficient to model the fossil data.



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

###Heteroscedasticity analysis

### Do another set of gam cubic spline fit of residuals to get the fitted trend line. Do it by Fitted values

### and predictor, Age

### Using default value of K

### Fit for all choices of K = 4 , 8 , 23 and the fit will b eused later on

fit.gam.k4.abs\_resids = gam(abs(residuals(fit.gam.k4)) ~ s(Age , bs = "cr"))

fit.gam.k8.abs\_resids = gam(abs(residuals(fit.gam.k8)) ~ s(Age , bs = "cr"))

fit.gam.k23.abs\_resids = gam(abs(residuals(fit.gam.k23)) ~ s(Age , bs = "cr"))

fit.gam.k4.abs\_resids.fitted = gam(abs(residuals(fit.gam.k4)) ~ s(fitted(fit.gam.k4) , bs = "cr"))

fit.gam.k8.abs\_resids.fitted = gam(abs(residuals(fit.gam.k8)) ~ s(fitted(fit.gam.k8) , bs = "cr"))

fit.gam.k23.abs\_resids.fitted = gam(abs(residuals(fit.gam.k23)) ~ s(fitted(fit.gam.k23) , bs = "cr"))

## Plot all the fits in ONE graph

dev.off()

par(mfrow = c(3 ,1))

plot(Age , Sr.Ratio , pch = "\*" , cex = 2 , col = 1 , main = "Fossil Data plot ")

####Main call of the plot. Plot the abs residuals against X and add a trend line

####for each fits i.e. fit with K = 4 , 8 , 23

####Plotting K = 4

plot(Age , abs(residuals(fit.gam.k4)) , pch = "\*" , cex = 2 , col = 4

, main = "Fossil Data plot - |Residual| Analysis by Age" , xlab = "Age" , ylab = "GAM |Residuals|" )

points(Age , fit.gam.k4.abs\_resids$fitted.values , col = 4 , type = "l" , lwd = 2)

####Plotting K = 8

points(Age , abs(residuals(fit.gam.k8)) , cex = 2, col = 5 , pch = "\*" , type = "p" )

points(Age , fit.gam.k8.abs\_resids$fitted.values , col = 5 , type = "l" , lwd = 2)

####Plotting K = 23

points(Age , abs(residuals(fit.gam.k23)) , cex = 2, col = 6 , pch = "\*" , type = "p" )

points(Age , fit.gam.k23.abs\_resids$fitted.values , col = 6 , type = "l" , lwd = 2)

legend("topright",

legend = c("|Residual| GAM K = 4" ,"|Residual| GAM K = 8","|Residual| GAM K = 23"),

col = 4:6,

cex = 1.1,

pch = "\*"

)

####Plotting K = 4 , residuals by fitted value

plot(fitted(fit.gam.k4) , abs(residuals(fit.gam.k4)) , pch = "\*" , cex = 2 , col = 4

, main = "Fossil Data plot - |Residual| Analysis by Fitted Values" , xlab = "Fitted values" , ylab = "GAM |Residuals|" )

points(fitted(fit.gam.k4) , fit.gam.k4.abs\_resids.fitted$fitted.values , col = 4 , type = "l" , lwd = 2)

####Plotting K = 8 , residuals by fitted value

points(fitted(fit.gam.k8) , abs(residuals(fit.gam.k8)) , cex = 2, col = 5 , pch = "\*" , type = "p" )

points(fitted(fit.gam.k8) , fit.gam.k8.abs\_resids.fitted$fitted.values , col = 5 , type = "l" , lwd = 2)

####Plotting K = 23 , residuals by fitted value

points(fitted(fit.gam.k23) , abs(residuals(fit.gam.k23)) , cex = 2, col = 6 , pch = "\*" , type = "p" )

points(fitted(fit.gam.k23) , fit.gam.k23.abs\_resids.fitted$fitted.values , col = 6 , type = "l" , lwd = 2)

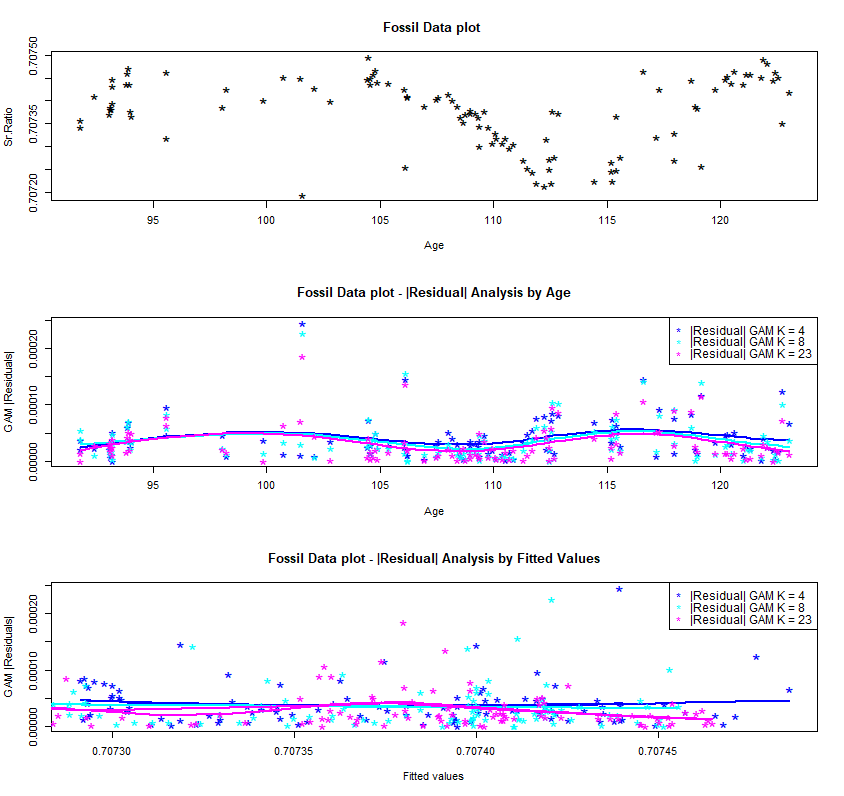
legend("topright",

legend = c("|Residual| GAM K = 4" ,"|Residual| GAM K = 8","|Residual| GAM K = 23"),

col = 4:6,

cex = 1.1,

pch = "\*"

)

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

####Ratio of maximum fitted absolute residual to minimum fitted absolute residual, for the

####check of C.I or the region and inference reliability

####K = 4

max.k4.fit.AbsRes = abs(residuals(fit.gam.k4))[which(fitted(fit.gam.k4) == max(fitted(fit.gam.k4)))]

min.k4.fit.AbsRes = abs(residuals(fit.gam.k4))[which(fitted(fit.gam.k4) == min(fitted(fit.gam.k4)))]

K4Ratio.Hetero = max.k4.fit.AbsRes / min.k4.fit.AbsRes

#K4Ratio.Hetero = .81

####K = 8

max.k8.fit.AbsRes = abs(residuals(fit.gam.k8))[which(fitted(fit.gam.k8) == max(fitted(fit.gam.k8)))]

min.k8.fit.AbsRes = abs(residuals(fit.gam.k8))[which(fitted(fit.gam.k8) == min(fitted(fit.gam.k8)))]

K8Ratio.Hetero = max.k8.fit.AbsRes / min.k8.fit.AbsRes

#K8Ratio.Hetero = .36

####K = 23

max.k23.fit.AbsRes = abs(residuals(fit.gam.k23))[which(fitted(fit.gam.k23) == max(fitted(fit.gam.k23)))]

min.k23.fit.AbsRes = abs(residuals(fit.gam.k23))[which(fitted(fit.gam.k23) == min(fitted(fit.gam.k23)))]

K23Ratio.Hetero =max.k23.fit.AbsRes / min.k23.fit.AbsRes

#K23Ratio.Hetero = .31

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

####

####We see, that none of the three fits with K = 4 , 8 and 23 violate the thumb rule the ratio > 3 and ratioSquared > 9.

####

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

####

####If the ratio of maximum fitted absolute residual to minimum fitted absolute residual is greater than 3

####or if the square of the same quantity is greater than 9 , then the confidence region for the fit to Y from

####X is not reliable.

####

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

####

####

####K is the number of Knots or interior Knots as in O' Sullivan splines. By specifying K knots, we mean

####that we believe that there are K changes (break points) on the predictor which changes the response variable.

####