###

####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 04

####

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

setwd("F:/OneDrive/Learning/DataScience/Statistics Texas A&M University/689/Assignment/Assignment3")

##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)

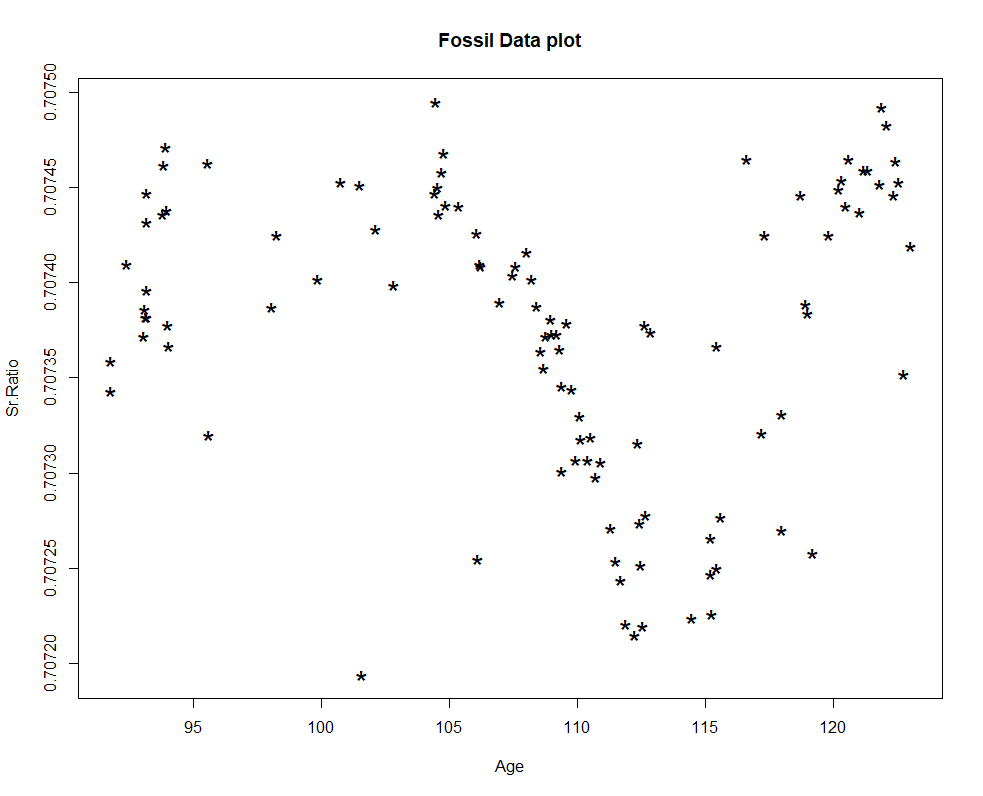
Age = age[order(age)]

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

###################Answers to Question 1###################

## 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 data

##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.

###################Answers to Question 2 a. and c. ###################

## Fit the fossil data using the default version of smooth.spline

## Reference the library from the HRW package

library("HRW")

##fit the model , keep options as default choices provided with the software

fit.smoothspline.fossil = smooth.spline(Age , Sr.Ratio)

##Draw the scatter plot as asked in 2.c

points((fit.smoothspline.fossil) , pch = "\*" , cex = 2, col = 2 , type = "b" , lwd = 2)

legend("bottomleft",

legend = c("Smooth spline fit" ),

col = 2,

cex = 1.1,

pch = "\*"

)

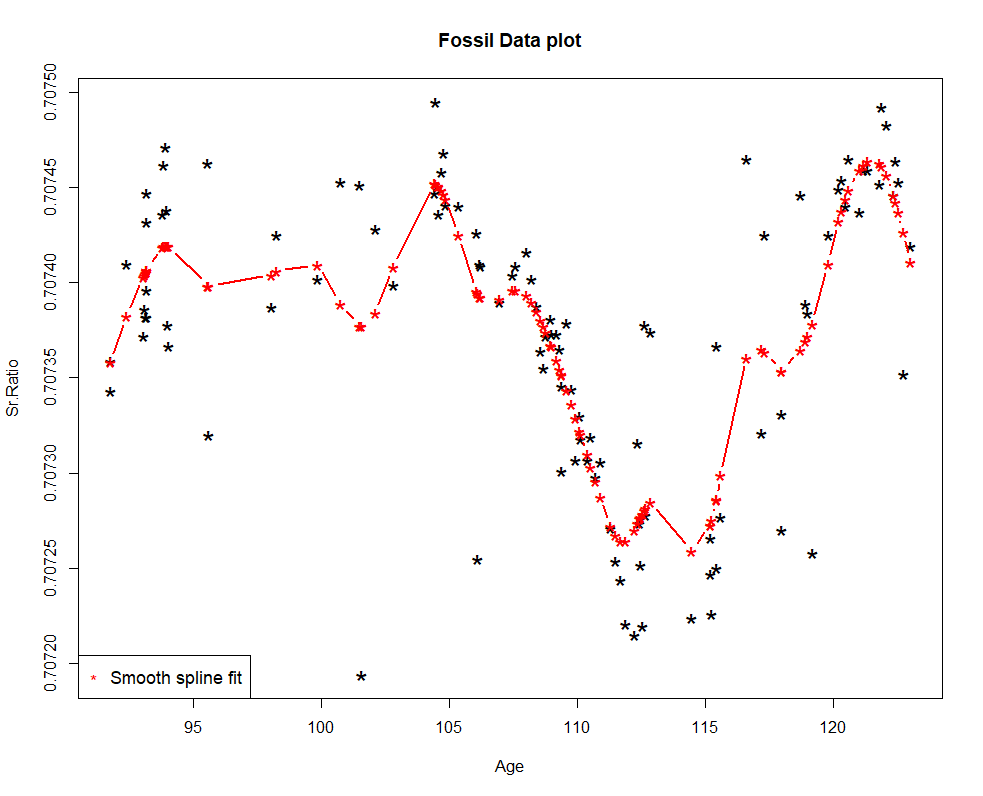
print(fit.smoothspline.fossil)

#Smoothing Parameter spar= 0.7031912 lambda= 1.368289e-05 (13 iterations)

#Equivalent Degrees of Freedom (Df): 17.86713

#Penalized Criterion (RSS): 1.996244e-07

#GCV: 2.725977e-09



###################Answers to Question 3 ###################

##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)

####Part a.

## 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.

####Part b.

## 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",

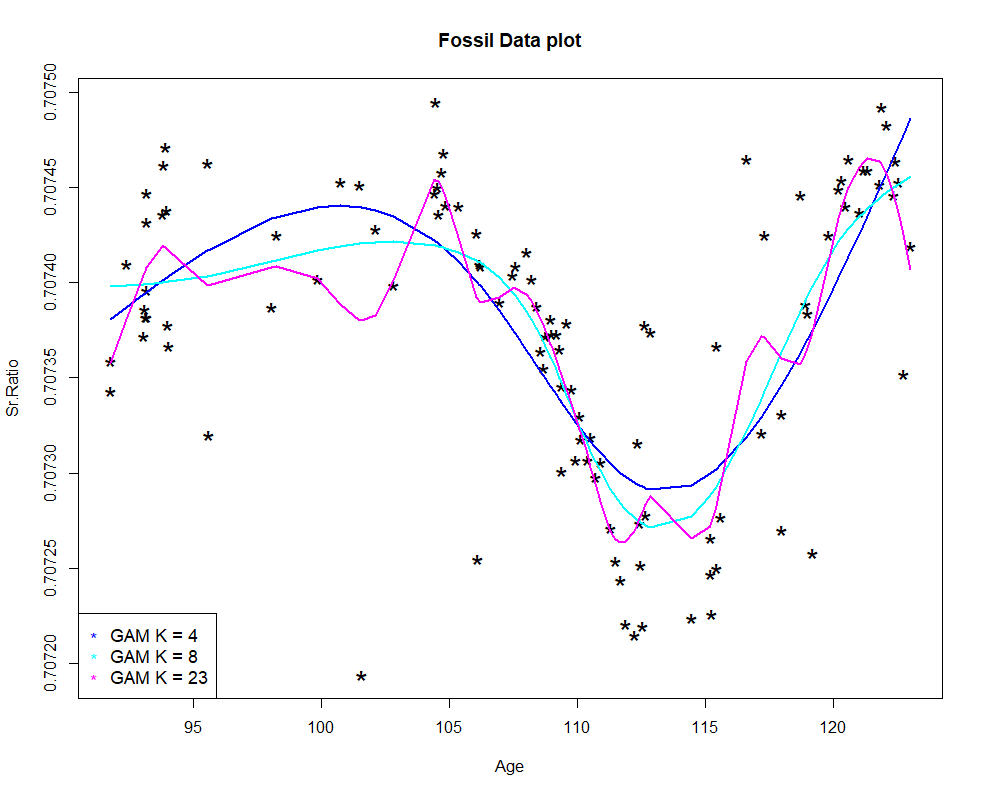
legend = c("GAM K = 4" ,"GAM K = 8","GAM K = 23"),

col = 4:6,

cex = 1.1,

pch = "\*"

)



####Part c.

#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.

###################Answers to Question 4 ###################

##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

###################Answers to Question 5 and 6 ###################

##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

###################Answers to Question 6 ###################

# 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.