Homework 4

Skylar Liu

2024-02-03

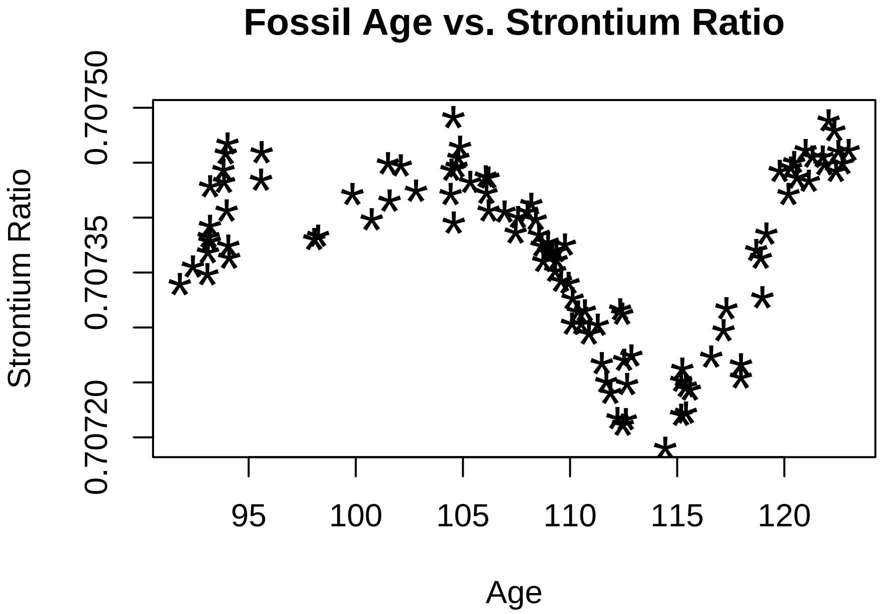
Homework #4, Stat 660, Spring 2024, Class #6, Monday February 5, 2024

1. Display and study the scatterplot of these data. What features of the data look interesting to you? Do this before answering the other questions.

```r
# clear workspace
rm(list = ls())
# set the seed
set.seed(382957)

# import data
fossil = read.csv("~/660 - Flexible Regression/Homework/Homework3/fossil.csv")
X = fossil$Age
Y = fossil$Strontium.Ratio

# scatterplot X vs Y
plot(X, Y, type="p",pch='*',main="Fossil Age vs. Strontium Ratio",
     xlab="Age",ylab="Strontium Ratio",cex=2)
```



The pattern of the data is interesting to me, particularly how the Strontium Ratio dips around 114 years of Age. It appears to be in a general cubic shape.

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

   a. Get and save the model object. You use something like myspline = smooth.spline(...). The model object is myspline.

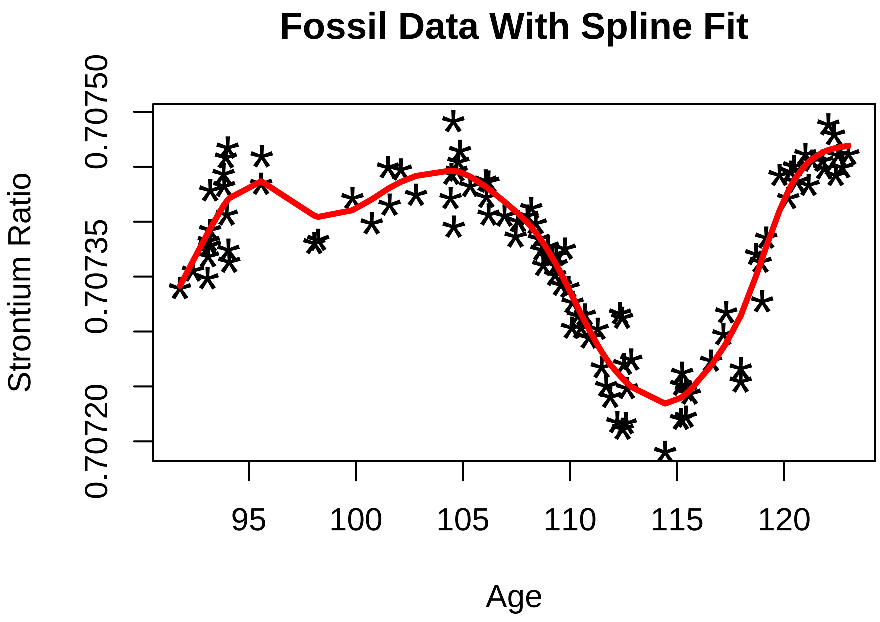
```
# Get the smooth.spline function
library(HRW)

# fit the smooth.spline model
myspline = smooth.spline(X, Y)
```

   b. Add the fitted line to the scatter plot of the data and display the resulting plot.

```
# use the predict function to setup the line
mypred = predict(myspline)

# plot the scatter plot and fitted line
plot(X, Y, type="p",pch='*',main="Fossil Data With Spline Fit",
     xlab="Age",ylab="Strontium Ratio",cex=2)
lines(mypred$x,mypred$y,lwd=3,col="red")
```



3. Run the mgcv fit to the data with the default number of knots (K=8) and with both K = 4 and K = 23 knots and using the cubic spline option as I have done. Save the model fit objects, e.g., gam4, gam8 and gam23.

   a. Which fits are statistically significant? Be sure to quote the p-values for all three.

```r
# Get the mgcv package
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.9-1. For overview type ’help("mgcv-package")’.
```

```r
# mgcv fit with K = 4 knots
gam4 = gam(Y~s(X,bs="cr",k=4))
summary(gam4)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Y ~ s(X, bs = "cr", k = 4)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.074e-01  3.350e-06  211142   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##        edf Ref.df     F p-value
## s(X) 2.998      3 135.9  <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.794   Deviance explained =   80%
## GCV = 1.2364e-09  Scale est. = 1.1898e-09  n = 106
```

```r
# mgcv fit with K = 8 knots
gam8 = gam(Y~s(X,bs="cr",k=8))
summary(gam8)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Y ~ s(X, bs = "cr", k = 8)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.074e-01  2.701e-06  261903   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
```

```
##        edf Ref.df     F p-value
## s(X) 6.502  6.907 98.09  <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.866   Deviance explained = 87.5%
## GCV = 8.3215e-10  Scale est. = 7.7325e-10  n = 106
```

```
# mgcv fit with K = 23 knots
gam23 = gam(Y~s(X,bs="cr",k=23))
summary(gam23)
```

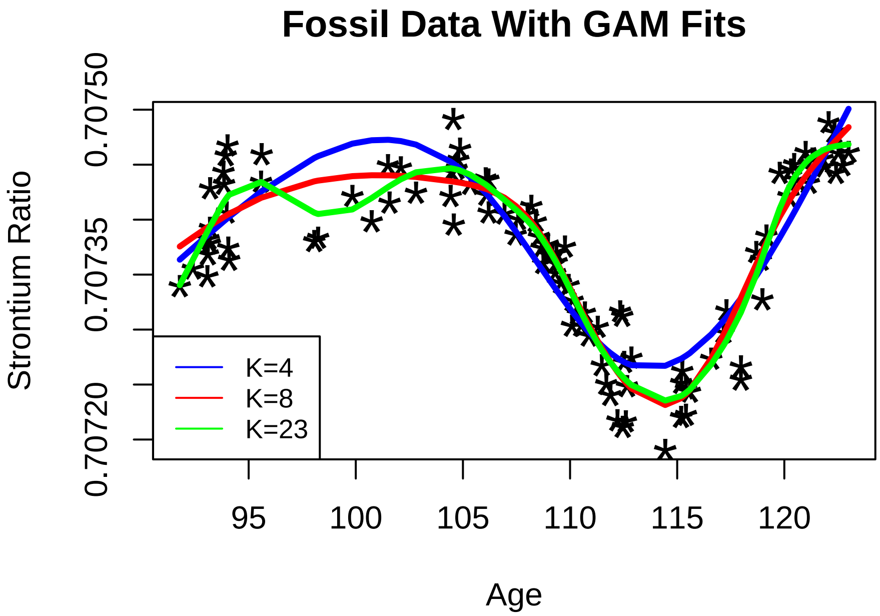
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Y ~ s(X, bs = "cr", k = 23)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.074e-01  2.427e-06  291474   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##        edf Ref.df     F p-value
## s(X) 11.52  13.76 63.11  <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.892   Deviance explained = 90.4%
## GCV = 7.079e-10  Scale est. = 6.2432e-10  n = 106
```

The p-value for all three knot values (k = 4, 8, and 23) is 2^-16. Since this value is less than alpha=.05, they are all statistically significant.

b. Plot the fits with the data points on one graph ONLY and submit that graph.

```
# sort the data
ord = sort(X, index.return = T)$ix

# plot the scatter plot and fitted lines
plot(X, Y, type="p",pch='*',main="Fossil Data With GAM Fits",
     xlab="Age",ylab="Strontium Ratio",cex=2)
lines(X[ord], fitted(gam4)[ord], lwd=3, col="blue")
lines(X[ord], fitted(gam8)[ord], lwd=3, col="red")
lines(X[ord], fitted(gam23)[ord], lwd=3, col="green")
legend("bottomleft", legend = c("K=4", "K=8", "K=23"), col=c("blue", "red", "green"),lty=1, cex=0.8)
```

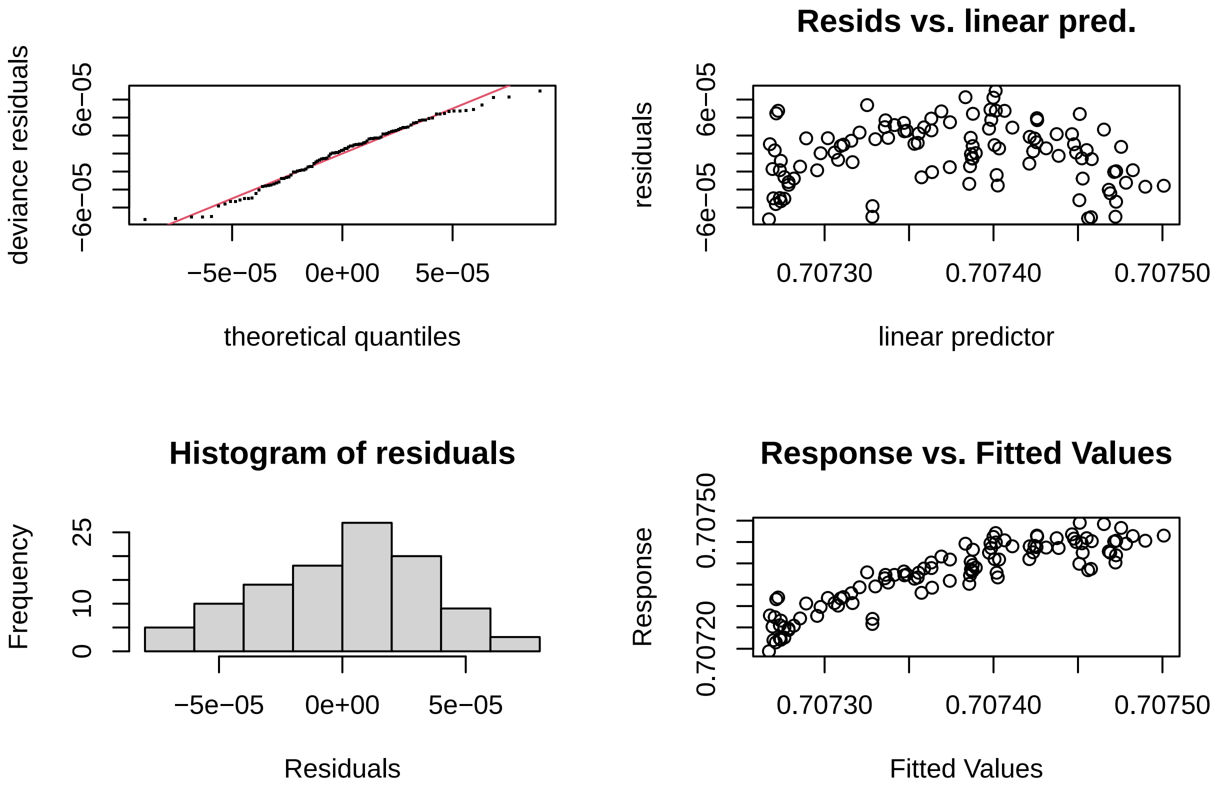


c. Do the fits agree more or less with your answer to Question 1? Why or why not?

In general, the plots create a cubic fit consistent with my observations. The K=23 fit deviates slightly, with bumps around age 100.

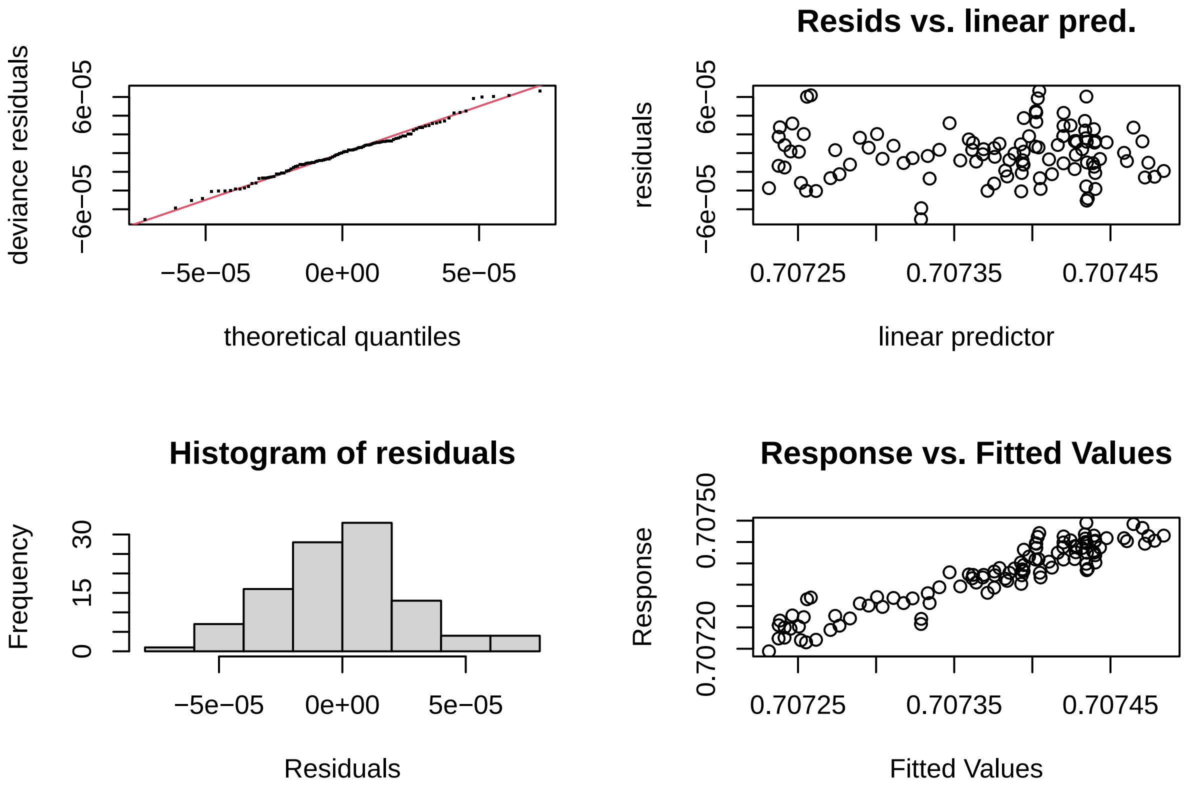
4. What are the effective degrees of freedom for each mgcv fit?

```
# check the effective degrees of freedom for each mgcv fit
gam.check(gam4)
```



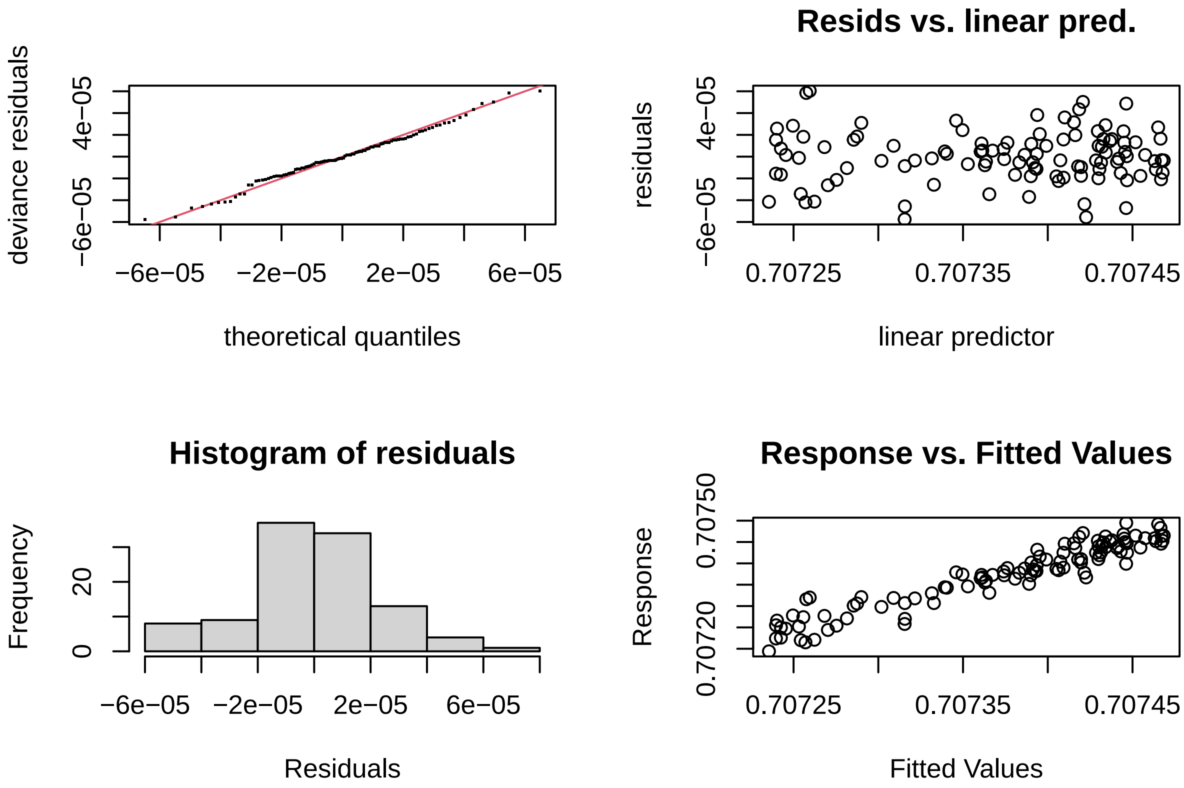
```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 4 iterations.
## The RMS GCV score gradient at convergence was 1.072308e-10 .
## The Hessian was positive definite.
## Model rank =  4 / 4
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##      k' edf k-index p-value
## s(X)  3   3    0.55  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gam.check(gam8)
```



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 4 iterations.
## The RMS GCV score gradient at convergence was 1.095501e-11 .
## The Hessian was positive definite.
## Model rank =  8 / 8
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##       k' edf k-index p-value
## s(X) 7.0 6.5    0.85    0.07 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gam.check(gam23)
```



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 4 iterations.
## The RMS GCV score gradient at convergence was 1.839388e-12 .
## The Hessian was positive definite.
## Model rank =  23 / 23
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##        k'  edf k-index p-value
## s(X) 22.0 11.5    1.05    0.68
```

EDF of gam4: 3 EDF of gam8: 6.5 EDF of gam23: 11.5

```
# lambda of each mgcv fit
gam4$sp
```

```
##       s(X)
## 0.01212764
```

```
gam8$sp
```

```
##      s(X)
## 0.5889037
```

```
gam23$sp
```

```
##      s(X)
## 7.493726
```

gam4: lambda = 0.01212764 gam8: lambda = 0.5889037 gam23: lambda = 7.493726

6. Tell me whether or not the p-value for each choice of K is < 0.10. Cite those p-values. If any are < 0.10, then explain intuitively from your graphs why that number of basis functions is inadequate.

   K = 4: p-value = <2e-16 < 0.10 K = 8: p-value = 0.07 < 0.10 K = 23: p-value = 0.68 < 0.10

   K=4 and K=8 p-values are < 0.10, suggesting an inadequate number of knots. These number of basis functions fail to pick up the bump that is present in K=23, around age 100.