Assignment 04 - STAT 689

Philip Anderson, panders2@tamu.edu

February 7, 2018

```
# Import third-party modules
library("mgcv")
```

Import our data and take a look:

```
fossil <- read.csv("C:/Users/Philip/Schools/TAMU/STAT_689/homework/semiparametric-regression/misc/fossil
names(fossil) <- tolower(names(fossil))
head(fossil)
```

```
##        age strontium.ratio
## 1 91.78525        0.707343
## 2 91.78525        0.707359
## 3 92.39579        0.707410
## 4 93.97061        0.707438
## 5 95.57577        0.707463
## 6 95.60286        0.707320
```

```
summary(fossil)
```

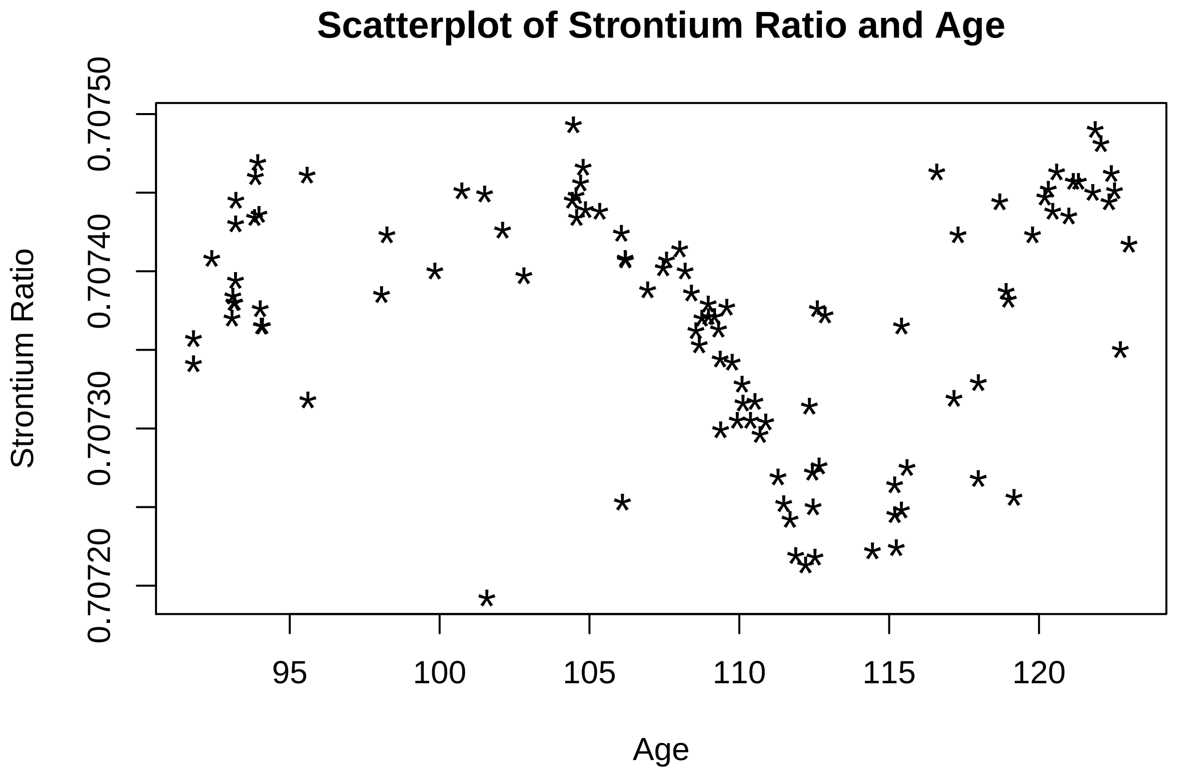
```
##       age          strontium.ratio
##  Min.   : 91.79   Min.   :0.7072
##  1st Qu.:103.62   1st Qu.:0.7073
##  Median :109.37   Median :0.7074
##  Mean   :108.62   Mean   :0.7074
##  3rd Qu.:115.41   3rd Qu.:0.7074
##  Max.   :123.00   Max.   :0.7075
```

Question 1

Display a scatterplot of the data and comment on which features appear interesting.

```
# non-generalizable function for this data/plot only
fossil_plotter <- function() {
plot(fossil$age, fossil$strontium.ratio, pch="*", cex=1.5
     , main="Scatterplot of Strontium Ratio and Age"
     , xlab="Age"
     , ylab="Strontium Ratio"
     )
    }

fossil_plotter()
```



The features of this data that appear most interesting are:

- The clear non-linear relationship between age and Strontium ratio.
- Some evidence of hetereoscedacticity
- Strontium ratio appears to have a higher variance at lower values of age.
- The necessity for a three-part piecewise regression, with a positive line from 90-105, negative from 105-112.5, and positive again from 112.5 to 125.

Question 2

Question 2a

Fit the fossil data using the default version of stats::smooth.spline

```
myspline <- stats::smooth.spline(x=fossil$age, y=fossil$strontium.ratio)
myspline
```

```
## Call:
## stats::smooth.spline(x = fossil$age, y = fossil$strontium.ratio)
##
## Smoothing Parameter  spar= 0.7240878  lambda= 1.216781e-05 (12 iterations)
## Equivalent Degrees of Freedom (Df): 18.38485
## Penalized Criterion (RSS): 1.992173e-07
## GCV: 2.716276e-09
```

Question 2B

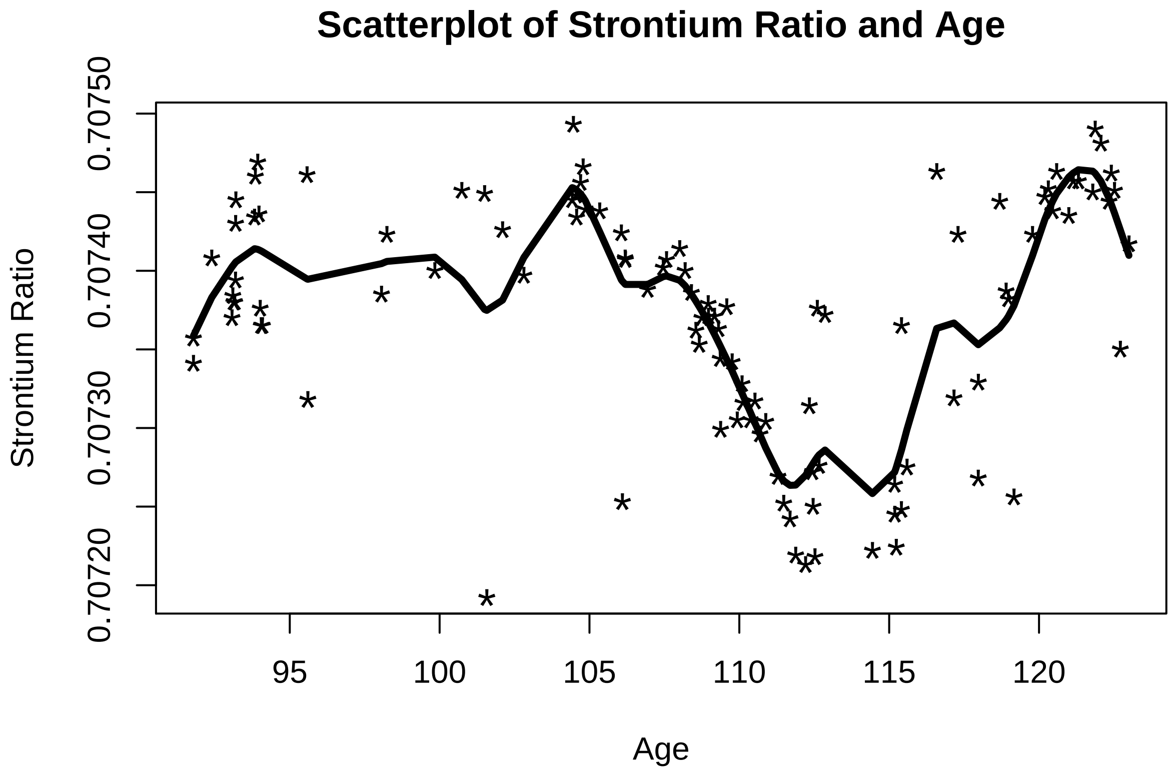
Is the fit statistically significant?

TBD

Question 2C

Add the fit to the scatterplot

```
fossil_plotter()
lines(myspline, lwd=3.5)
```



Question 3

Run the mgcv fit to the data with the default number of knots (8), along with 4 and 23 using the cubic spline option.

Using a simple function to save space.

```
fossil_gam <- function(knots=8) {
  gam_mod <- mgcv::gam(strontium.ratio ~ s(age, k=knots, bs="cr")
                       , data=fossil
                       )
  return(gam_mod)
```

```r
}

gam8 <- fossil_gam(knots=8)
gam4 <- fossil_gam(knots=4)
gam23 <- fossil_gam(knots=23)
```

Question 3a

Which fits are statistically significant?

I am using gam.check(.) to arrive at these values. I am not displaying the gam.check(.) output, as it renders many plots, and the plotting functionality does not appear to be disable-able. It is also not clear from the gam.check soure code how to just extract p-values.

```r
paste0("GAM with 8 knots spline p-value: 0.46")
```

```
## [1] "GAM with 8 knots spline p-value: 0.46"
```

```r
paste0("GAM with 4 knots spline p-value: 0.08")
```

```
## [1] "GAM with 4 knots spline p-value: 0.08"
```

```r
paste0("GAM with 23 knots spline p-value: 0.97")
```

```
## [1] "GAM with 23 knots spline p-value: 0.97"
```

Only the GAM fit with 4 knots has a statistically significant slope.

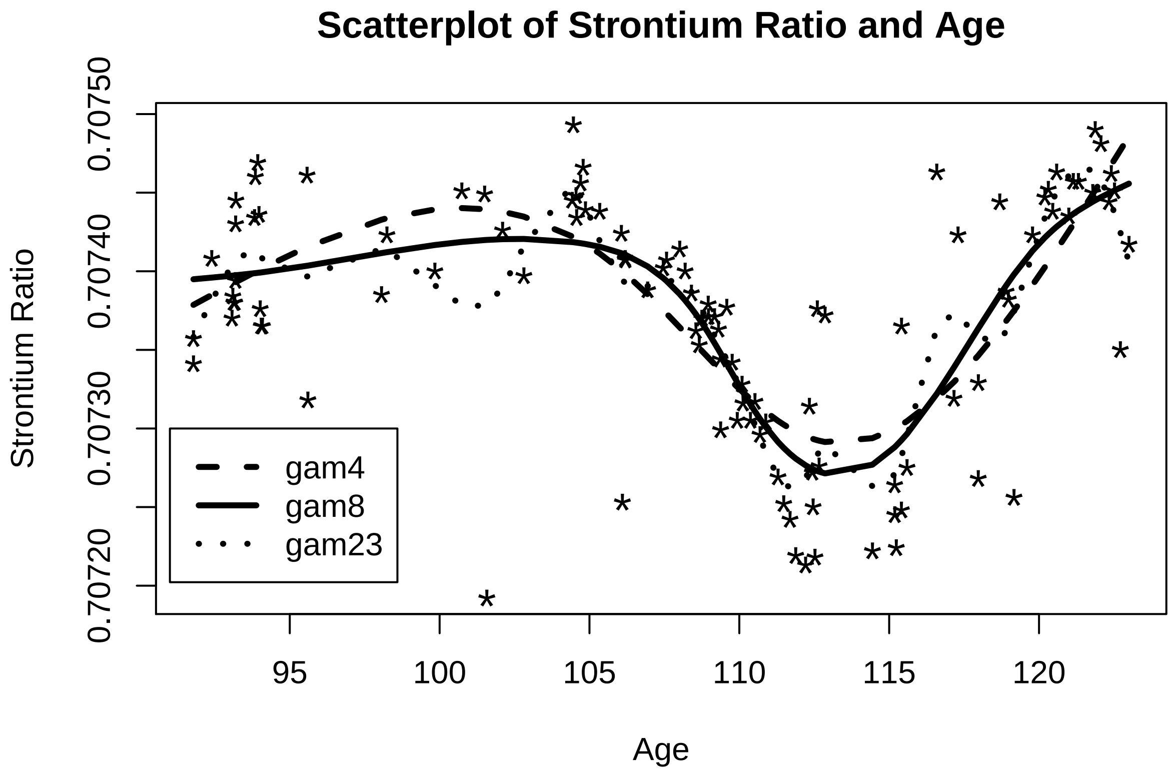
Question 3b

Plot the data points on one graph only.

```r
# point ordering for plots
age_ord <- order(fossil$age)

# initiate plots
fossil_plotter()
lines(fossil$age[age_ord], gam8$fitted.values[age_ord], lty=1, lwd=3)
lines(fossil$age[age_ord], gam4$fitted.values[age_ord], lty=2, lwd=3)
lines(fossil$age[age_ord], gam23$fitted.values[age_ord], lty=3, lwd=3)

legend(91, 0.7073
       , c("gam4", "gam8", "gam23")
       , lty=c(2, 1, 3)
       , lwd= rep(3,3)
       )
```



Question 3c

The fits from all gam models are consistent with my initial observation of non-linearity. The smoothing parameters for all fits also break the regression into three main pieces, with an ascending piece followed by a descending piece, again followed by an ascending piece.

Question 4

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

```
paste0("GAM with 8 knots edf: ", sum(gam8$edf) - 1) # subtract 1 for one predictor model
```

```
## [1] "GAM with 8 knots edf: 5.99019961739269"
```

```
paste0("GAM with 4 knots edf: ", sum(gam4$edf) - 1)
```

```
## [1] "GAM with 4 knots edf: 2.98781801014555"
```

```
paste0("GAM with 23 knots edf: ", sum(gam23$edf) - 1)
```

```
## [1] "GAM with 23 knots edf: 15.8526035231863"
```

Question 5

What is lambda for each fit?

```
paste0("GAM with 8 knots lambda value: " , gam8$sp)
```

```
## [1] "GAM with 8 knots lambda value: 1.59552509993438"
```

```
paste0("GAM with 4 knots lambda value: " , gam4$sp)
```

```
## [1] "GAM with 4 knots lambda value: 0.0990383701373478"
```

```
paste0("GAM with 23 knots lambda value: " , gam23$sp)
```

```
## [1] "GAM with 23 knots lambda value: 1.19545806847315"
```

Question 6

Are any of the p-values less than 0.10?

With a p-value of 0.08, only the GAM with 4 knots has an inadequate fit. This is likely because the data displays a number of small trend fluctuations that only 4 knots are unable to capture. We can see in the most recent graphic that the GAM with 4 knots is unable to capture a necessary amonunt of the variation present in our data.