STAT 689-600 HW04

Dennis Leet

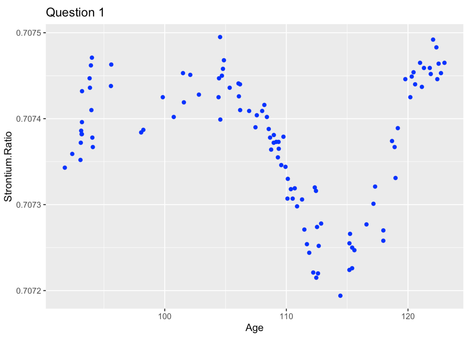
2/5/2022

Question 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.

Answer

There are three main movements in the data approximately occurring over the following intervals: 0 to 105, 105 to 115 and 115 and greater. Heteroskedasticity is also apparent throughout the domain. As referenced in the previous exercise the data generating function appears to be cubic.

```
#Simple ggplot:
ggplot(data = fossil)+geom_point(aes(x=Age, y=Strontium.Ratio), col="blue") + ggtitle("Question 1")
```



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

Answer

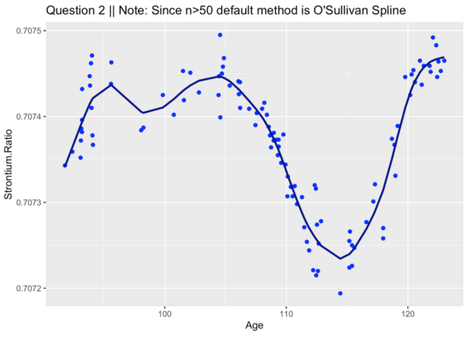
a. Get and save the model object. You use something like myspline = smooth.spline(...). The model object is myspline.

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

```
#fitting a default smooth.spline() from the 'stats' package.
#lambda/nknots parameters are automatically selected. Default for
#n>50 is O'Sullivan due to computational efficiency.
fossilSpline <- smooth.spline(x,y)

#Predicting and overlaying default spline on scatte
fossilSplinePred <- predict(fossilSpline,newdata=as.data.frame(x),se.fit=TRUE)

#Plotting
ggplot(data = fossil)+geom_point(aes(x=Age, y=Strontium.Ratio), col="blue") + ggtitle("Question 2 || Note: Since
n>50 default method is O'Sullivan Spline")+
geom_line(aes(x = fossilSplinePred$x, y = fossilSplinePred$y), col="darkblue", size=1)
```



Question 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.
b. Plot the fits with the data points on one graph ONLY and submit that graph.
c. Do the fits agree more or less with your answer to Question 1? Why or why not?

Answer

a. All fits appear to be *highly* significant with all **p-values <2e-16** on s(Age) at k = 4, 8 and 23. Similarly, all levels of K carry significant intercepts.

b. See below for plot.

c. Yes. Each model captures the three major movements occurring in the data. Since they are all incredibly significant it is difficult to say which is best. I can see how you could easily burn your entire career up searching for an optimal number of knots! From a purely visual perspective I am drawn to the model with K = 4. It seems to capture the important trends across the domain while being more resistant to noise in the sample. The model with k=23 is too reactive and probably distorts the *cubicesque* function which likely generated the data.

```
##MGCV Package Fit. Creating models with cubic spline (bs=cr) models
# with k = 4,8 and 23.

#k=4
gamK4 <- gam(y~s(x, k=4, bs="cr"), data=fossil)
#summary(gamK4)

#k=8 or default
gamK8 <- gam(y~s(x,bs="cr"), data=fossil)
#summary(gamK8)

#k=23
gamK23 <- gam(y~s(x, k=23,bs="cr"), data=fossil)
#summary(gamK23)

#One Plot for our 3 GAM Models.
ggplot(data = fossil)+geom_point(aes(x=Age, y=Strontium.Ratio), col="black") + ggtitle("Q3.b: Splines || Maroon:
K = 4 || Gold: K = 8 || Purple: K = 23")+
geom_line(aes(Age, y = fitted(gamK4)), col="maroon", size=1) +
geom_line(aes(Age, y = fitted(gamK8)), col="gold", size=1) +
geom_line(aes(Age, y = fitted(gamK23)), col="purple", size=1)
```


```
#We now seek to assess the overall model fit. Note that
#lower p-values often correspond to less desirable fits.
#gam.check(gamK4)
#gam.check(gamK8)
#gam.check(gamK23)
l <- matrix(c("K","4","8","23","lambda = ",as.vector(gamK4$sp),as.vector(gamK8$sp),as.vector(gamK23$sp)), nrow =
4, byrow = F)
l
```

```
##      [,1] [,2]
## [1,] "K"  "lambda = "
## [2,] "4"  "0.0121276423469896"
## [3,] "8"  "1.82326428584335"
## [4,] "23" "7.49372567289754"
```

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

Answer

The Effective Degrees of freedom are 3, 8.06 and 11.5 at K = 4, 8 and 23. There is an interesting message in the **gam.check()** output indicating that K might be too small if the p-value is close to zero *and* K' and the EDF are close numerically. This appears to be the case at K = 4. *This ends up being important in question 6.*

Question 5 What is lambda for each fit?

Answer

If **K = (4, 8, 23)** *then* **lambda = (0.012, 1.823, 7.494)** respectively.

Question 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.

Answer

The p-values are <2e-16 at K=4, 0.28 at K=8 and 0.68 at K=23. The process occurring in the gam.check() GCV is randomized so these p-values change each time they are used - *on the same data!* A heuristic exploration reveals that while they occur randomly, they seem to be distributed so variably as to appear on either side of the cut off value easily. However, there are always degenerate cases.

I will admit that my knee jerk reaction would have been the K=4 model. Retrospectively, I can see that at least k=8 seems to be more fitting specifically on the domain interval close to Age = 100. On these values of Age the k = 4 model far overestimates the conditional mean of the strontium levels. Also around Age = 133, the K = 4 fit with 2 spline basis function doesn't dip low enough to capture the signal. The other two functions, with their larger number of basis functions, differ somewhat for lower values of Age and then approximate each other for Age > 105. The driving point here seems to be that at K=4, there simply aren't enough spline basis functions to capture the signal.