Assignment 05 - STAT 689

Philip Anderson: panders2@tamu.edu

February 7, 2018

```r
# import third-party modules
library("mgcv")
```

Import our data

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

Fit GAM's with 4, 8, and 23 knots

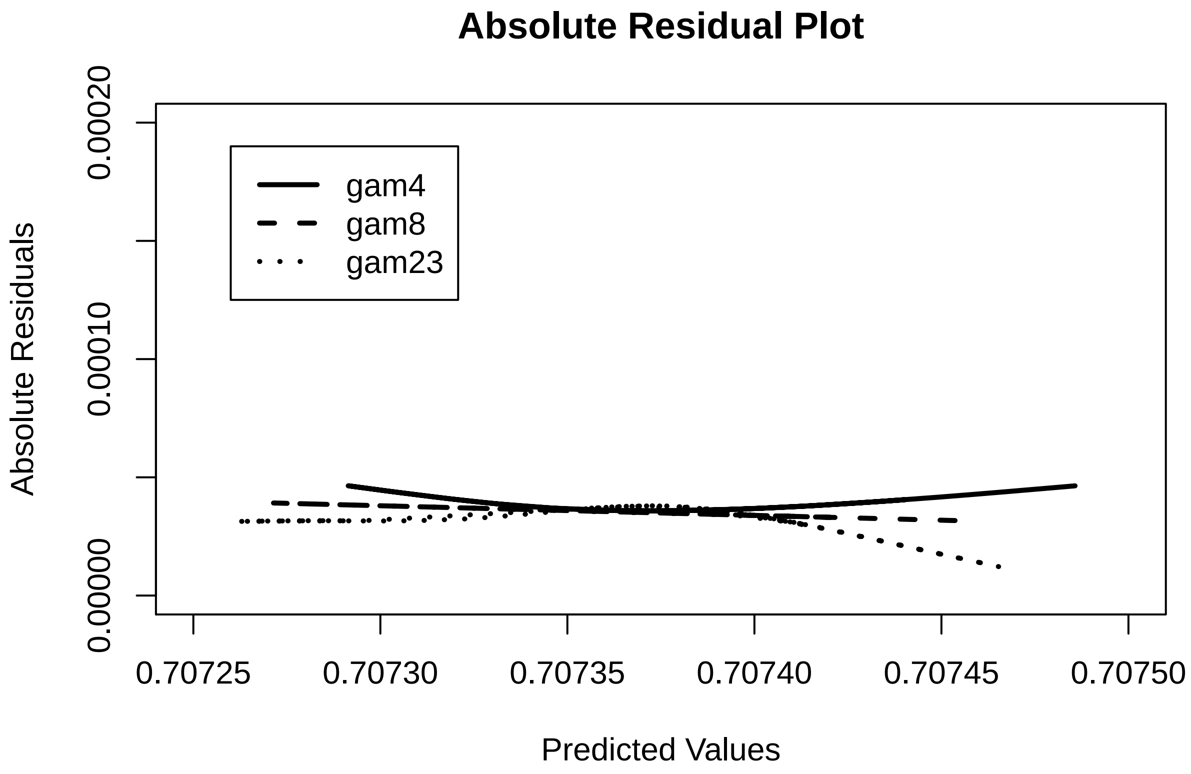
```r
fossil_mod <- function(knots=8) {
  mod <- mgcv::gam(strontium.ratio ~ s(age, k=knots, bs="cr")
                   , data=fossil )
  return(mod)
              }

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

Question 1

For each of the three models, plot the absolute fitted residuals against the predicted values.

```r
plot(1, type="n"
     , xlim=c(0.70725, 0.70750)
     , ylim=c(0, 0.00020)
     , xlab="Predicted Values"
     , ylab="Absolute Residuals"
     , main="Absolute Residual Plot"
     )

line_plotter <- function(gam_obj, lty=1, knots=8){
      age_ord <- order(fossil$age)
      x <- fitted(gam_obj)[age_ord]
      y <- abs(gam_obj$residuals)[age_ord]
      gam_obj_abs <- mgcv::gam(y ~ s(x, k=knots, bs="cr"))
      lines(x, fitted(gam_obj_abs), lwd=2.5, lty=lty)
      }
```

```r
line_plotter(gam_obj=gam4, lty=1, knots=4)
line_plotter(gam_obj=gam8, lty=2, knots=8)
line_plotter(gam_obj=gam23, lty=3, knots=23)

legend(0.70726, 0.00019
       , c("gam4", "gam8", "gam23")
       , lty=c(1, 2, 3)
       , lwd= rep(2.5,3)
       )
```



Question 2

Give the ratio of the maximum fitted absolute residual to the minimum absolute fitted residual.

```r
ratio_finder <- function(mod_obj, precision=2) {
  # find the ratio of the max value of abs(resid) to the min value
     ratio <- round(max(abs(mod_obj$residuals)) / min(abs(mod_obj$residuals)), precision)
     return(ratio)
}

paste0("GAM with 4 knots ratio is: ", ratio_finder(mod_obj=gam4))
```

```
## [1] "GAM with 4 knots ratio is: 89.82"
```

```
paste0("GAM with 8 knots ratio is: ", ratio_finder(mod_obj=gam8))
```

```
## [1] "GAM with 8 knots ratio is: 771.81"
```

```
paste0("GAM with 23 knots ratio is: ", ratio_finder(mod_obj=gam23))
```

```
## [1] "GAM with 23 knots ratio is: 320.81"
```

Question 3

All three of the ratios dramatically exceed our rule of thumb value of 3.

```
data.frame(K=c(4,8,23)
          , Ratio=c(ratio_finder(mod_obj=gam4)
                    , ratio_finder(mod_obj=gam8)
                    , ratio_finder(mod_obj=gam23)
                    )
           )
```

```
##    K  Ratio
## 1  4  89.82
## 2  8 771.81
## 3 23 320.81
```

Question 4

The failure of our results to fall within the bounds of the residual ratio heuristic indicates that our model fails to capture the non-constant variance present in our single predictor variable. It indicates that our model is widely missing in making at least one prediction, which is the consequence of heteroscedacticity. I have included more extensive residual plots in the Appendix of this document, which more thoroughly demonstrate the problem.

Question 5

Give a verbal description of what K is.

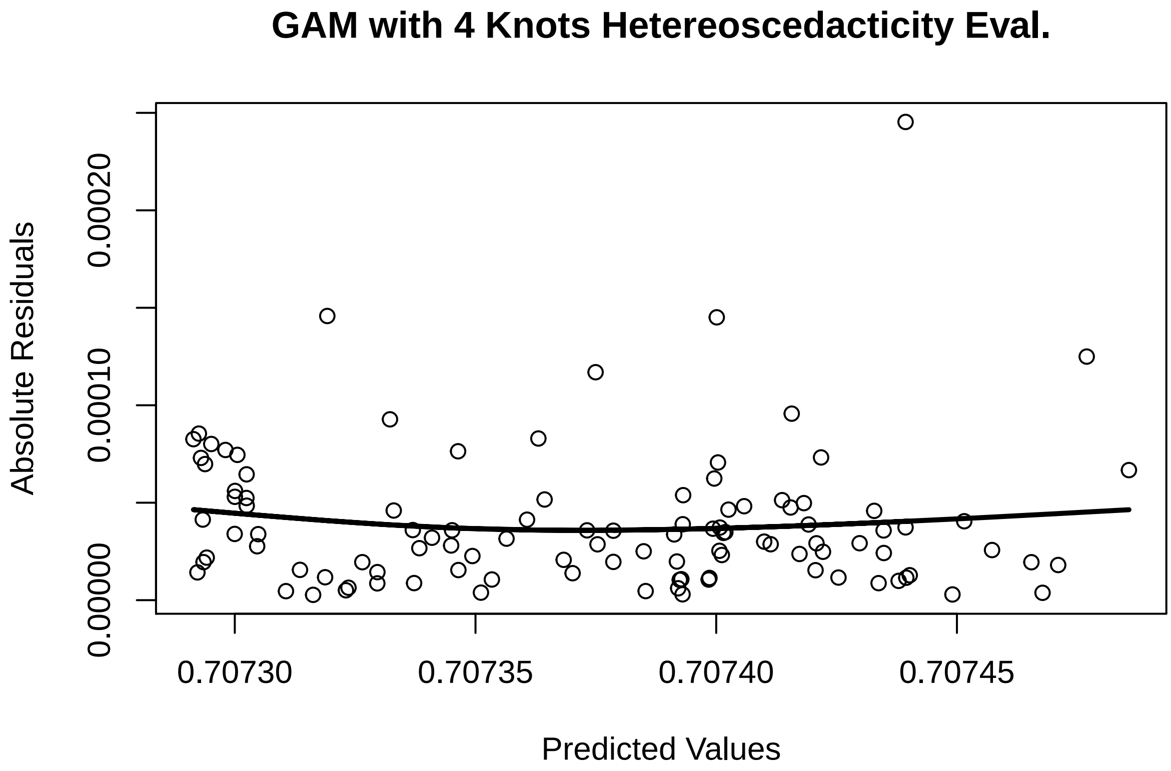
K is the number of basis dimensions, or knots, with which our smoothed regression line will be fit. These are the points at which our fitted line will be permitted to deviate from linearity. It is selected through identifying the unique quantiles of the x variable, with a forumula such as:

$$knot\ location = (\frac{k}{K-1})th\ quantile\ of\ unique\ x$$

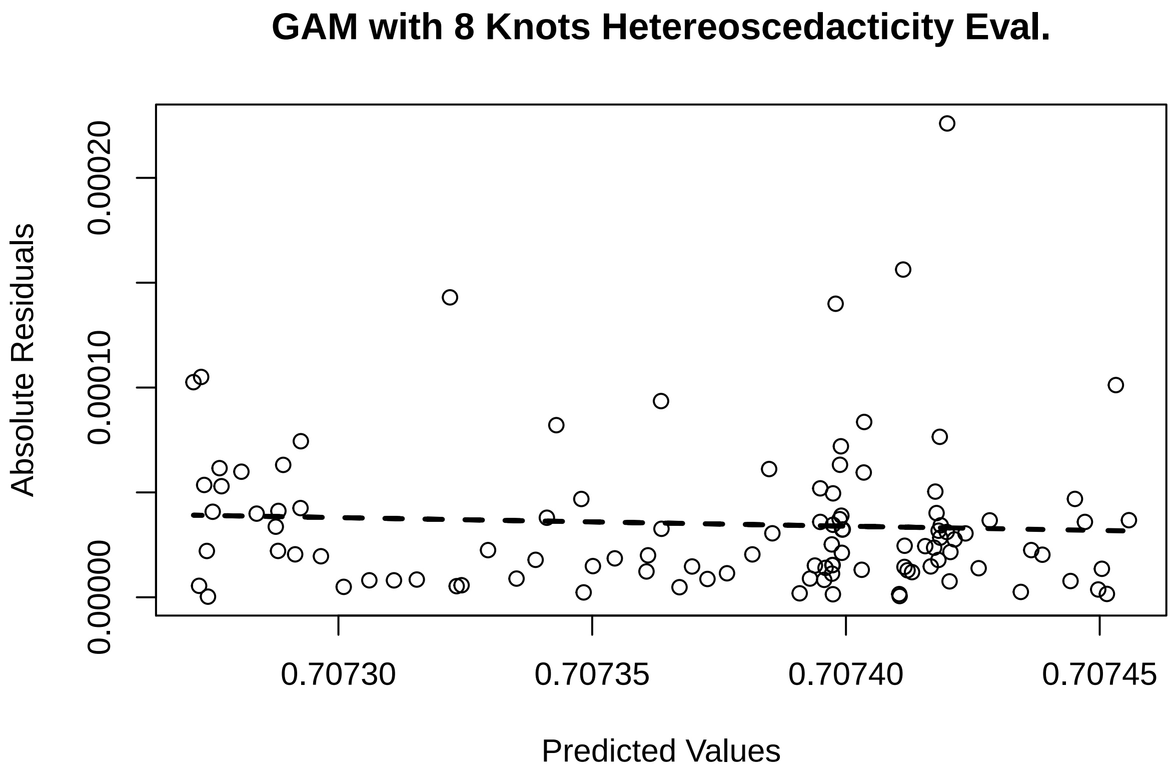
Appendix

Question 1 expanded. I include distinct hetereoscedacticity plots for each of the fitted models (4, 8, and 23 knots).

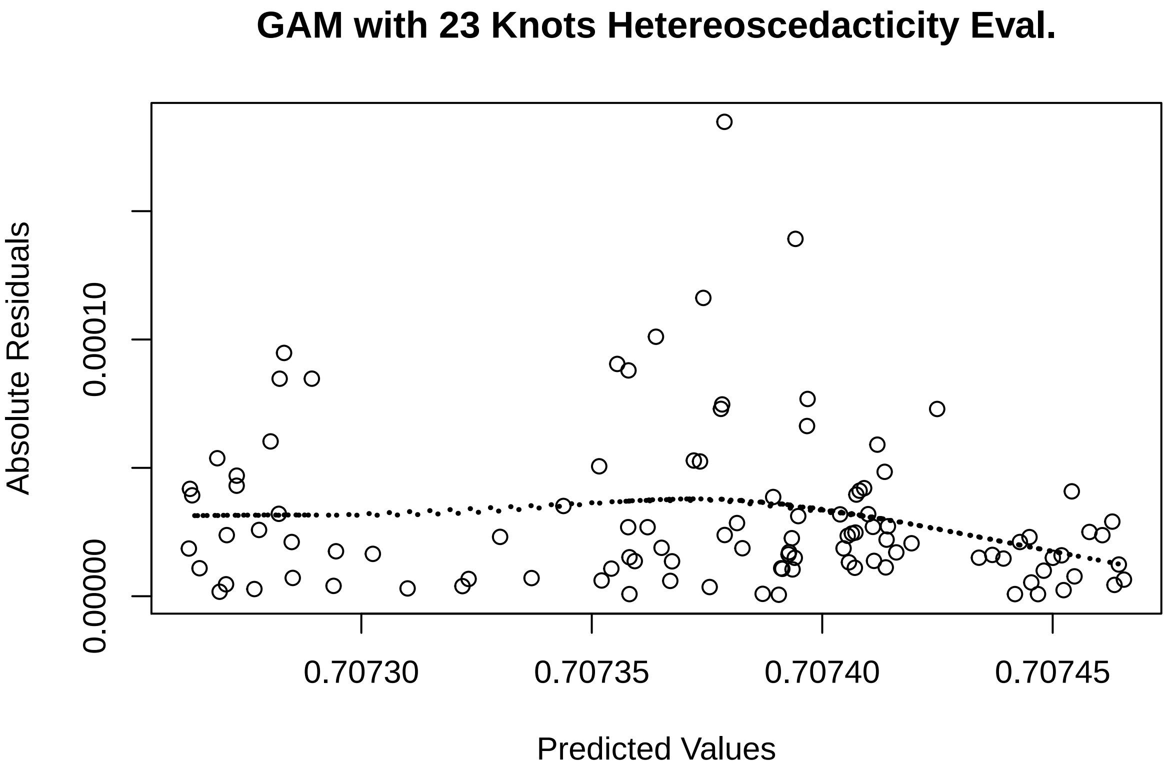
```r
hs_plotter <- function(gam_obj, knots, title_prompt="Fitted Vals vs. Abs. Resid.", lty=1) {
        # create a scatterplot and put a smooth regression on it
        plot(fitted(gam_obj), abs(gam_obj$residuals)
             , xlab="Predicted Values"
             , ylab="Absolute Residuals"
             , main=title_prompt
             )

        age_ord <- order(fossil$age)
        x <- fitted(gam_obj)[age_ord]
        y <- abs(gam_obj$residuals)[age_ord]
        gam_obj_abs <- mgcv::gam(y ~ s(x, k=knots, bs="cr"))
        lines(x, fitted(gam_obj_abs), lwd=2.5, lty=lty)


}

hs_plotter(gam_obj=gam4, knots=4, title_prompt="GAM with 4 Knots Hetereoscedacticity Eval.", lty=1)
```



```
hs_plotter(gam_obj=gam8, knots=8, title_prompt="GAM with 8 Knots Hetereoscedacticity Eval.", lty=2)
```



```
hs_plotter(gam_obj=gam23, knots=23, title_prompt="GAM with 23 Knots Hetereoscedacticity Eval.", lty=3)
```


GAM with 23 Knots Hetereoscedacticity Eval.
Absolute Residuals
0.00000
0.00010
0.70730
0.70735
0.70740
0.70745
Predicted Values