BST210_HW_3

Hillary Miller

September 20, 2017

BST 210 HOMEWORK #3

Due 8:00 AM, Monday, September 25, 2016

Here we continue to explore data from the â??Singapore Cardiovascular Cohort Study 2â??, using continuous age, gender, and continuous body mass index (defined as weight/height2 in kg/m2) to predict total cholesterol (in mmol/l) of subjects. Note that 1 mmol/l (SI units) equals 38.67 mg/dl, the usual American units for cholesterol. Our main goal is to assess potential nonlinear effects of continuous covariates through the use of appropriate spline or GAM models (you have some flexibility here and only need to use one of these approaches [your choice, which may depend on your favorite statistical package] to answer the questions below).

1. First, we further explore the effects of continuous age to predict total cholesterol.

(a) Run three linear regression models using linear age, linear and quadratic age, and then either a spline or GAM modeling of age to predict total cholesterol. In one sentence, describe how you have fitted the spline or GAM model (e.g., choice of knot points, order of the polynomial, other restrictions). Plot the three sets of fitted values, look at the regression output obtained, and briefly compare the three fits. Which of the three models do you recommend as being â??bestâ?? so far? Why?

I'm fitting the following three models:

$$E[tc] = \beta_0 + \beta_1(age)$$

$$E[tc] = \beta_0 + \beta_1(age) + \beta_2(age^2)$$

$$E[tc] = \beta_0 + f(age)$$

```
library(knitr)
knitr::opts_chunk$set(echo=TRUE)
library(haven)
library(ggplot2)
library(MASS)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(car)

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

library(splines2)
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

library(mgcv)

## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-21. For overview type 'help("mgcv-package")'.

homeworkdata <-read_dta("C:/Users/millerhillaryv/Desktop/HSPH/BST210/BST 210
homework/HW3/homeworkdata.dta")
homework <- homeworkdata
set.seed(210)
```

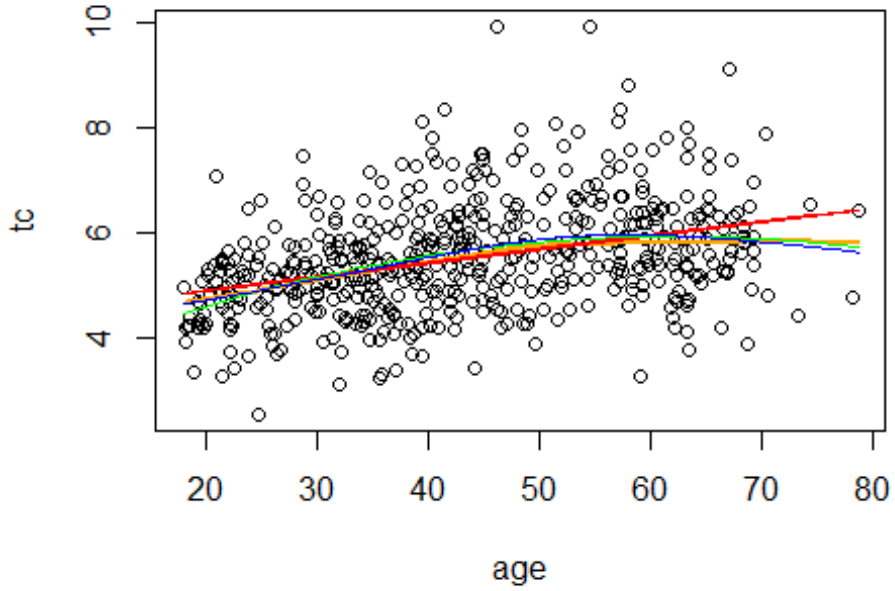
```
homework$bmi<-homework$weight/((homework$height)/100)^2
homework$BMICat<-with(homework,
ifelse(homework$bmi < 18.5, "2",
ifelse(homework$bmi < 25, "1",
ifelse(homework$bmi <30, "3", "4"))))

homework$bmi_sq <-homework$bmi^2

##_males <-filter(homework)

homework$age_sq <- homework$age^2

number1 <- data.frame("tc"=homework$tc, "age"=homework$age, "age2"=homework$a
ge_sq)
number1[sort(number1$age),]




modlm=lm(tc ~ age, data=number1)
modquad=lm(tc ~ age + age2, data=number1)
modgam=gam(tc~s(age,k=4,bs="cr"), data=number1)

summary(modlm)

##
## Call:
## lm(formula = tc ~ age, data = number1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6589 -0.6383 -0.0557  0.5009  4.3216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.376191   0.135848  32.214   <2e-16 ***
## age         0.026243   0.002966   8.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9801 on 525 degrees of freedom
## Multiple R-squared:  0.1298, Adjusted R-squared:  0.1281
## F-statistic: 78.31 on 1 and 525 DF,  p-value: < 2.2e-16

summary(modquad)

##
## Call:
## lm(formula = tc ~ age + age2, data = number1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -2.6542 -0.6410 -0.0461  0.5151  4.1698
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.0658041  0.3907005   7.847 2.41e-14 ***
## age          0.0920305  0.0186508   4.934 1.08e-06 ***
## age2        -0.0007389  0.0002069  -3.572 0.000387 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9693 on 524 degrees of freedom
## Multiple R-squared:  0.1505, Adjusted R-squared:  0.1472
## F-statistic: 46.41 on 2 and 524 DF,  p-value: < 2.2e-16
```

```
summary(modgam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## tc ~ s(age, k = 4, bs = "cr")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51742    0.04214   130.9   <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(age) 2.619  2.895 31.73  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.15   Deviance explained = 15.5%
## GCV = 0.94251  Scale est. = 0.93604   n = 527
```

```
plot(tc~age, data=number1)
lines(lowess(number1$tc~number1$age), col="orange", lwd=2)
lines(number1$age,fitted(modlm),col='red')
lm.fit2 = lm(tc ~ poly(age, 2, raw = TRUE), data = number1)
curve(predict(lm.fit2, newdata = data.frame(age = x)), add = TRUE, col = "gre
en")
curve.number1 = data.frame(x=number1$age, y=predict(modgam))
curve.number1 = curve.number1[order(curve.number1$x),]

lines(curve.number1, col="blue")
```



I fit a GAM model by choosing 4 knot points with the cubic regression splines based on the output of the loess curve. Based on the plot of the three fitted values, the GAM model and model with linear and quadratic age are very similar to the loess curve. Evaluating the adjusted r-squared values of each model:

```
summary(modlm)$r.sq

## [1] 0.1297966

summary(modquad)$r.sq

## [1] 0.1504794

summary(modgam)$r.sq

## [1] 0.1504653
```

I would conclude that the 'best' model so far is the model of linear and quadratic age. This is due to the adjusted r-squared of the linear model being very similar to the gam model. While techincally, the gam model fits the data slightly better, the linear model is more parismonious and it is easier to interpret.

(b) Possibly you could confirm whether or not the linear age model is nested within your spline or GAM model by comparing your spline or GAM model to the model that also adds in linear age to your spline or GAM model. What happens when you do that? Can you tell if the linear age model is nested within your spline or GAM model? Briefly, how?

We are comparing the two models:

$$E[tc] = \beta_0 + GAM(age)$$

$$E[tc] = \beta_0 + \beta_1(age) + GAM(age)$$

We are testing whether the difference of sum of squares between the two models if very small. If so, then age adds nothing to the model, and age is nested in the GAM(age) model.

```
modgamage =gam(tc~s(age,k=4)+age, data=number1)
summary(modgam)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## tc ~ s(age, k = 4, bs = "cr")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51742    0.04214   130.9   <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(age) 2.619  2.895 31.73  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.15   Deviance explained = 15.5%
## GCV = 0.94251  Scale est. = 0.93604   n = 527

summary(modgamage)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## tc ~ s(age, k = 4) + age
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.50793    0.05105   9.949   <2e-16 ***
```

```
## age          0.11520    0.00146  78.909   <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(age) 2.514  2.792 362.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 4/5
## R-sq.(adj) =   0.15   Deviance explained = 15.5%
## GCV = 0.94262  Scale est. = 0.93615   n = 527
```

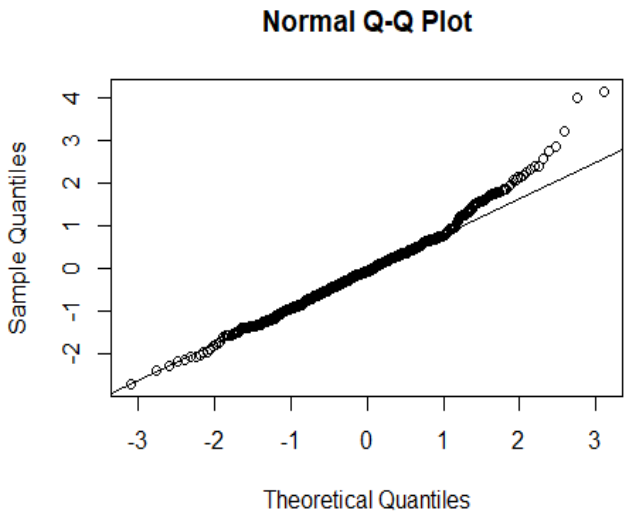
```
summary(modgam)$r.sq
```

```
## [1] 0.1504653
```

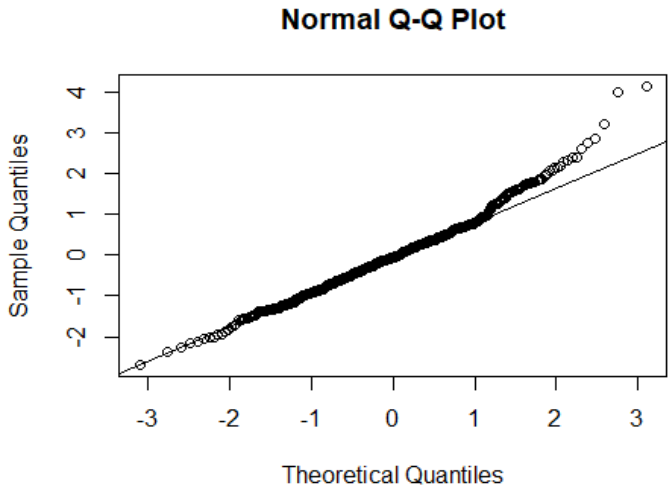
```
summary(modgamage)$r.sq
```

```
## [1] 0.1503576
```

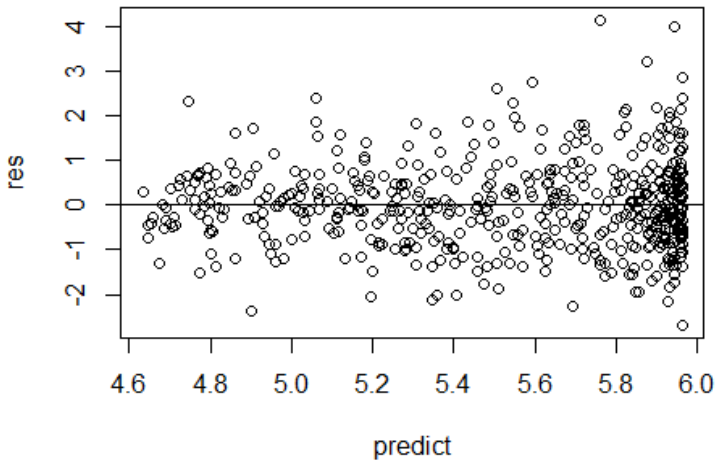
```
res1 <-residuals(modgamage)
res <- residuals(modgam)
predict <-predict(modgamage)
predict1 <-predict(modgam)

qqnorm(res1)
qqline(res1)
```



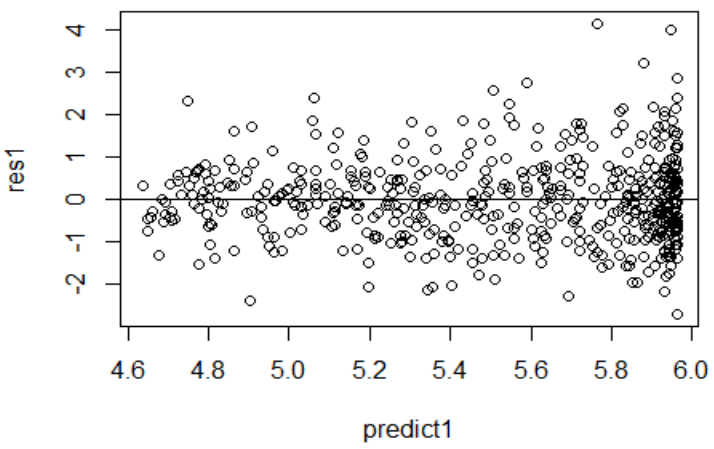

```
qqnorm(res)
qqline(res)
```




```r
plot(res~predict)
abline(h=0)
```




```r
plot(res1~predict1)
abline(h=0)
```



Looking at the plots of the residuals, both the linear+gam and the gam models appear normal, with little difference in the residuals plots of the two models. There is some heteroscedasticity as age increases.

```
sum(residuals(modgam)^2)

## [1] 489.9031

sum(residuals(modgamage)^2)

## [1] 489.971

anova(modgam,modgamage)

## Analysis of Deviance Table
##
## Model 1: tc ~ s(age, k = 4, bs = "cr")
## Model 2: tc ~ s(age, k = 4) + age
##   Resid. Df Resid. Dev         Df  Deviance
## 1    523.10     489.90
## 2    523.11     489.97 -0.0039936 -0.067873
```

When adding linear age into the model, the adjusted r-squared value remains almost constant. When evaluating the difference in the sum of the squares, the result is close to zero (0.068), and thus we can conclude that linear age is not adding anything to the model, and is therefore nested in the gam model.

(c) Can you tell if the linear and quadratic age model is nested within your spline or GAM model? Briefly, what are your findings?

To test if the modquad is nested inside modgam, I will run the following models:

$$E[tc] = \beta_0 + GAM(age)$$

$$E[tc] = \beta_0 + \beta_1(age) + \beta_2(age^2) + GAM(age)$$

```
modgamage2 =gam(tc~s(age,k=4)+age+age2, data=number1)
summary(modgamage2)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## tc ~ s(age, k = 4) + age + age2
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3649888  0.1205811   3.027  0.00259 **
## age          0.1530022  0.0323156   4.735 2.83e-06 ***
## age2        -0.0007154  0.0006161  -1.161  0.24608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df     F  p-value
## s(age) 1.526   1.84 29.48 1.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 5/6
## R-sq.(adj) =   0.15   Deviance explained = 15.4%
## GCV = 0.94326  Scale est. = 0.93677   n = 527

summary(modgam)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## tc ~ s(age, k = 4, bs = "cr")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.51742    0.04214   130.9   <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(age) 2.619  2.895 31.73  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

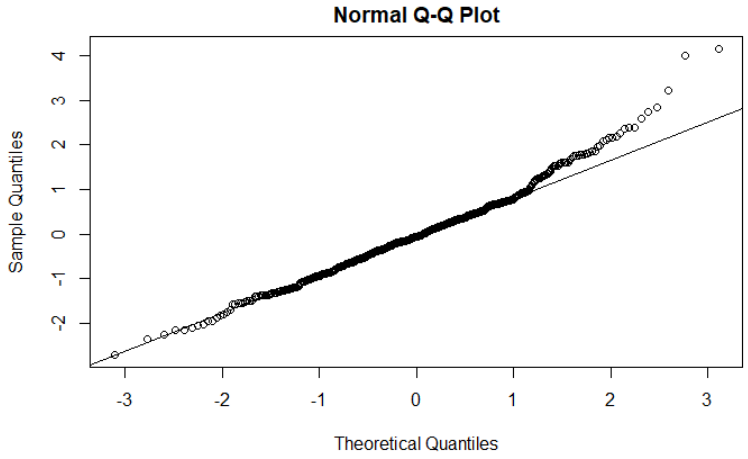
```
## R-sq.(adj) =   0.15   Deviance explained = 15.5%
## GCV = 0.94251  Scale est. = 0.93604   n = 527
```

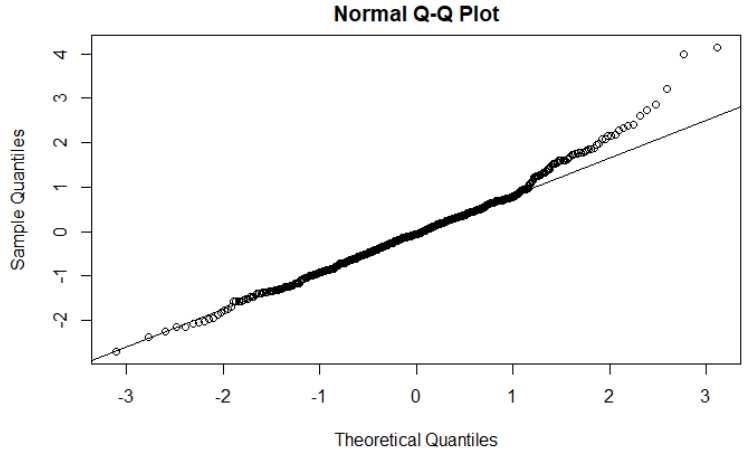
```
summary(modgamage2)$r.sq
```

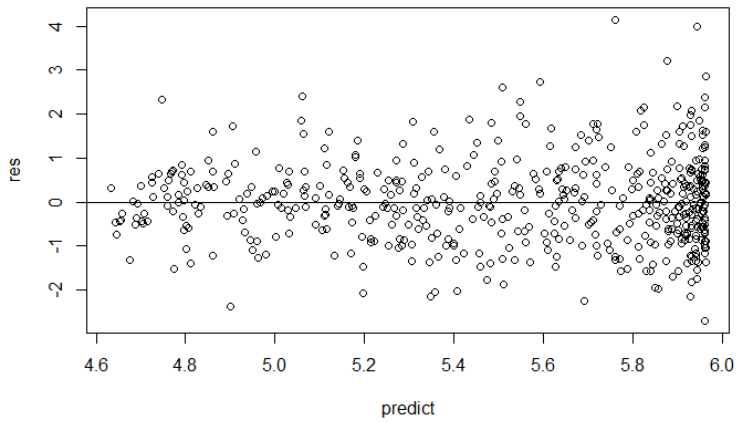
```
## [1] 0.1497954
```

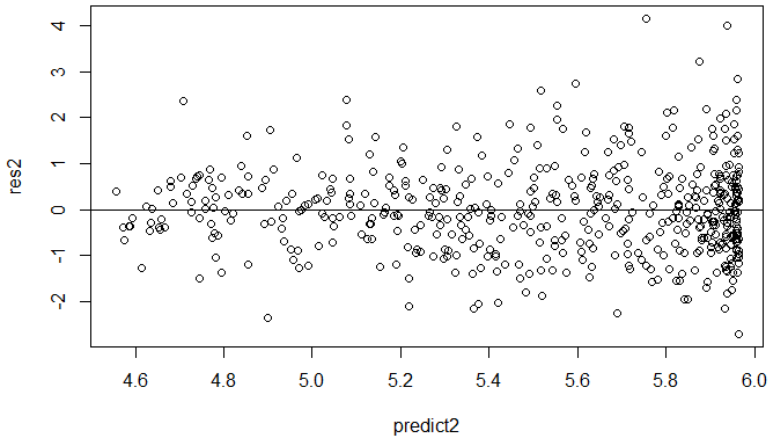
```{r}
res2 <-residuals(modgamage2)

predict2 <-predict(modgamage2)
```

```{r}
qqnorm(res2)

qqline(res2)

qqnorm(res)

qqline(res)

plot(res~predict)

abline(h=0)

plot(res2~predict2)

abline(h=0)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3
-2
-1
0
1
2
3
4


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-3
-2
-1
0
1
2
3
-2
-1
0
1
2
3
4


res
predict
4.6
4.8
5.0
5.2
5.4
5.6
5.8
6.0
-2
-1
0
1
2
3
4



Looking at the plots of the residuals, both the models appear normal, with little difference in the qqplots for the two models. Looking at the residual plots, the models suggest some heterscasticity as the values of age increase.

```
sum(residuals(modgam)^2)

## [1] 489.9031

sum(residuals(modgamage2)^2)

## [1] 490.2843

anova(modgam,modgamage2)

## Analysis of Deviance Table
##
## Model 1: tc ~ s(age, k = 4, bs = "cr")
## Model 2: tc ~ s(age, k = 4) + age + age2
##   Resid. Df Resid. Dev       Df Deviance
## 1    523.10     489.90
## 2    523.06     490.28 0.043596 -0.38117
```

When testing the semiparametric model with the addition of quadratic age, it can be concluded that quadratic age is nested in the GAM model. Looking at the difference of the sum of the squared residuals, the result is again close to 0, indicating that age^2 does not contribute much more to the model. Additionally, we can compare the edf of the original GAM to the one with linear and quadratic age. The original contained an edf of 2.69, indicating something between a quadratic and cubic fit. However, once quadratic age was added, the edf decreased, but the nonparametric component did still provide additional information. Thus, looking at the difference of the sum of the squares and the degrees of freedom, age and age^2 are nested inside the GAM model.

(d) You can also run a model using linear and quadratic age plus either a spline or GAM modeling of age, and determine how/whether we can tell if using linear and quadratic age is sufficient to model the effects of age versus a more complex spline or GAM model. Effectively, you are asking whether or not the spline or GAM modeling of age is needed after including linear and quadratic age. What are your conclusions? Be sure to perform an appropriate hypothesis test with an appropriate number of degrees of freedom.

$$E[tc] = \beta_0 + \beta_1 age + \beta_2 (age^2)$$

$$E[tc] = \beta_0 + \beta_1 age + \beta_2 (age^2) + f(age))$$

Ho = _3(GAM(age)) = 0 Ha = if _3(GAM(age)) !=0

```
modgam2 = gam(tc~age+ age2 + s(age,k=4), data=number1)
anova(modquad,modgam2)

## Analysis of Variance Table
##
## Model 1: tc ~ age + age2
## Model 2: tc ~ age + age2 + s(age, k = 4)
##   Res.Df    RSS      Df Sum of Sq      F  Pr(>F)
## 1 524.00 492.35
## 2 523.38 490.28 0.62444    2.0621 3.5252 0.07699 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

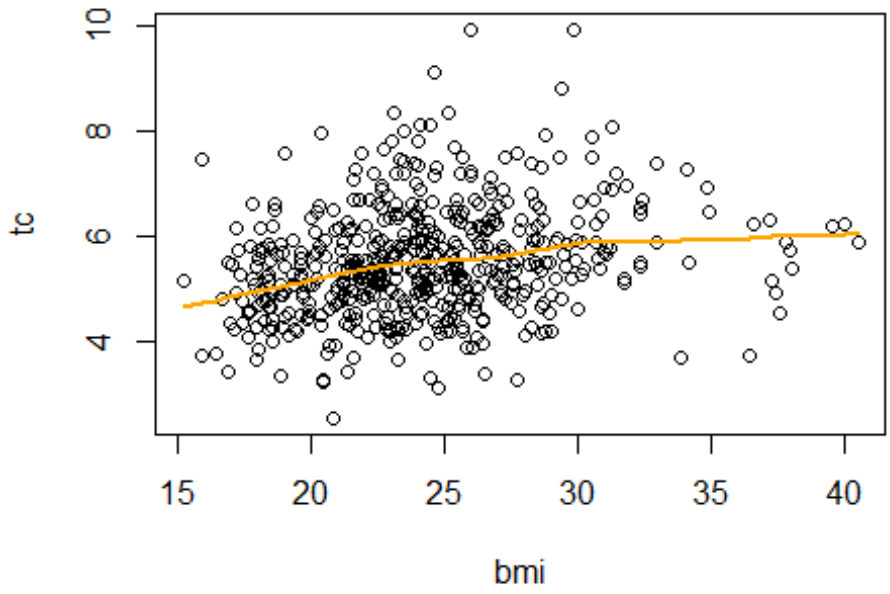
When testing the gam model with the addition of quadratic age to determine if the parametric model is sufficient, it can be concluded that the model including age and age^2 is sufficient for modeling the effects on age on tc. After conducting an F-test, the nonlinear effect of age is not statistically significant compared to the parametric model, with a p value of .06534, which is above the P<.05 threshold. Therefore, we cannot reject the null that GAM(age)=0 and cannot conlcude that adding a gam with f(age) contributes additional information to the model.

2. Suppose that the main research question is to determine the effects of (continuous) body mass index on total cholesterol, considering (continuous) age and gender as possible confounders or effect modifiers. The goal is to flexibly model the effects of age and gender while appropriately assessing the effects of body mass index on total cholesterol. You want to (hopefully!) be able to present an easily interpretable effect of body mass index on total cholesterol to your readers.

(a) Run some models that appropriately address this research question. What final model do you recommend? Briefly justify your choice. (You don't have to include the outputs of lots of models here, but perhaps write a brief description of your approach to get to your final model.)

Checking linearity of BMI and it's effect on cholesterol:

```
plot(tc~bmi, data=homework)
lines(lowess(homework$tc~homework$bmi), col="orange", lwd=2)
```



Since the relationship of BMI and cholesterol appears generally linear, will consider with additional potential confounders with three models:

1.Linear model with BMI, age (assuming we don't know information about the associationg between age and tc), and gender (assuming we don't know there is no association between gender and tc):

$$E[tc] = \beta_0 + \beta_1(bmi) + \beta_2(age) + \beta_3(gender)$$

2. Spline that addresses the complexity of age in the model:

$$E[tc] = \beta_0 + \beta_1(bmi) + \beta_2(spline(age)) + \beta_3(gender)$$

3. A GAM that allows for age to be evaluated in a nonparametric form:

$$E[tc] = \beta_0 + \beta_1(bmi) + \beta_2(GAM(age)) + \beta_3(gender)$$

Looking at the potential effects of BMI and gender: When fitting the model with gender as a covariate, the P value for the coefficient of gender is 0.5233 > 0.05, thus gender alone is not statistically significant. When fitting the model with age, the P value for the coefficient of age is <<.001, thus age is statistically significant, and could be a confounder. From the models below, we can see that adding the main effect of gender to the model of tc~bmi does not lead to any substantial change in the coefficient estimates of bmi. The slope for bmi has decreased from 0.05825 to 0.05806 (a 0.33% decrease). Since this is much less

than the 10% rule of thumb, gender does not appear to be a meaningful confounder on the effect of bmi on total cholesterol. Gender also does not appear to be an independent predictor of total cholesterol, as the p-value associated with its slope is well above the .05 threshold for statistical significance (p=.727).
We can also see that adding the main effect of age to the model of tc~bmi leads to the slope of bmi decreasing from 0.05825 to 0.0305255, which is above the 10% rule of thumb, concluding that age is a confounder of the effects of bmi on total cholesterol. Additionally, we see that the coefficient for the interaction term involving bmi is not statistically significant from zero (p=.49878), and therefore we conclude that gender does not act as an effect modifier of the relationship between linear BMI and total cholesterol.

```
modg <- lm(tc~gender, data=homework)
moda <-lm(tc~age, data=homework)
summary(modg)

##
## Call:
## lm(formula = tc ~ gender, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9508 -0.6908 -0.0908  0.6242  4.4392
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.49080    0.06189  88.722   <2e-16 ***
## gender       0.05870    0.09190   0.639    0.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 525 degrees of freedom
## Multiple R-squared:  0.0007765,  Adjusted R-squared:  -0.001127
## F-statistic: 0.408 on 1 and 525 DF,  p-value: 0.5233

summary(moda)

##
## Call:
## lm(formula = tc ~ age, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6589 -0.6383 -0.0557  0.5009  4.3216
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.376191   0.135848  32.214   <2e-16 ***
## age         0.026243   0.002966   8.849   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9801 on 525 degrees of freedom
## Multiple R-squared:  0.1298, Adjusted R-squared:  0.1281
## F-statistic: 78.31 on 1 and 525 DF,  p-value: < 2.2e-16
```

```
modb <-lm(tc~bmi, data=homework)
mod1 <-lm(tc~bmi+gender, data=homework)
mod2 <-lm(tc~bmi+age, data=homework)
summary(modb)
```

```
##
## Call:
## lm(formula = tc ~ bmi, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7806 -0.6333 -0.1273  0.5735  4.2889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.11005    0.25007  16.435  < 2e-16 ***
## bmi          0.05825    0.01019   5.719  1.8e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 525 degrees of freedom
## Multiple R-squared:  0.05864,    Adjusted R-squared:  0.05685
## F-statistic:  32.7 on 1 and 525 DF,  p-value: 1.803e-08
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = tc ~ bmi + gender, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7671 -0.6273 -0.1286  0.5823  4.2722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.10053    0.25176  16.288  < 2e-16 ***
## bmi          0.05806    0.01021   5.687 2.15e-08 ***
## gender       0.03126    0.08940   0.350    0.727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 524 degrees of freedom
## Multiple R-squared:  0.05886,    Adjusted R-squared:  0.05527
## F-statistic: 16.39 on 2 and 524 DF,  p-value: 1.251e-07
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = tc ~ bmi + age, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7346 -0.6345 -0.0704  0.5514  4.2665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.64997    0.24562  14.860  < 2e-16 ***
## bmi          0.03581    0.01014   3.532 0.000448 ***
## age          0.02305    0.00307   7.506 2.63e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9696 on 524 degrees of freedom
## Multiple R-squared:   0.15,  Adjusted R-squared:  0.1468
## F-statistic: 46.25 on 2 and 524 DF,  p-value: < 2.2e-16
```

```
modeffect <-lm(tc~bmi+age+gender+bmi:gender, data=homework)
summary(modeffect)
```

```
##
## Call:
## lm(formula = tc ~ bmi + age + gender + bmi:gender, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7255 -0.6474 -0.0820  0.5336  4.2259
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.424239   0.376780   9.088  < 2e-16 ***
## bmi          0.042914   0.015542   2.761  0.00596 **
## age          0.023384   0.003085   7.580 1.59e-13 ***
## gender       0.413954   0.485336   0.853  0.39409
## bmi:gender  -0.013394   0.019788  -0.677  0.49878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.97 on 522 degrees of freedom
## Multiple R-squared:  0.1526, Adjusted R-squared:  0.1461
## F-statistic:  23.5 on 4 and 522 DF,  p-value: < 2.2e-16
```

1. Checking a linear model of BMI and Cholesterol, adjusting for age and gender:

```
mod3<-lm(tc~bmi+age+gender, data=homework)
summary(mod3)
```

```
##
## Call:
## lm(formula = tc ~ bmi + age + gender, data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6952 -0.6376 -0.0776  0.5258  4.2178
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.616386   0.247619  14.605  < 2e-16 ***
## bmi         0.034961   0.010169   3.438 0.000633 ***
## age         0.023346   0.003083   7.573 1.66e-13 ***
## gender      0.090561   0.085312   1.062 0.288941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9695 on 523 degrees of freedom
## Multiple R-squared:  0.1519, Adjusted R-squared:  0.147
## F-statistic: 31.22 on 3 and 523 DF,  p-value: < 2.2e-16
```

2. checking a model of BMI and cholesterol that incorporates a piece-wise linear spline for age and also includes gender:

```
mod4=lm(tc~bmi+bSpline(age,df=4,degree=2)+gender, data=homework)
summary(mod4)

##
## Call:
## lm(formula = tc ~ bmi + bSpline(age, df = 4, degree = 2) + gender,
##     data = homework)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7459 -0.6042 -0.0793  0.5348  4.0964
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.99050    0.29965  13.317  < 2e-16 ***
## bmi                               0.02984    0.01025   2.911  0.00376 **
## bSpline(age, df = 4, degree = 2)1  0.25959    0.33569   0.773  0.43971
## bSpline(age, df = 4, degree = 2)2  0.91195    0.21844   4.175 3.50e-05 ***
## bSpline(age, df = 4, degree = 2)3  1.44051    0.32200   4.474 9.45e-06 ***
## bSpline(age, df = 4, degree = 2)4  0.62667    0.46196   1.357  0.17551
## gender                            0.07094    0.08503   0.834  0.40447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9624 on 520 degrees of freedom
## Multiple R-squared:  0.1689, Adjusted R-squared:  0.1593
## F-statistic: 17.61 on 6 and 520 DF,  p-value: < 2.2e-16
```

```
anova(mod3,mod4)
```

```
## Analysis of Variance Table
##
## Model 1: tc ~ bmi + age + gender
## Model 2: tc ~ bmi + bSpline(age, df = 4, degree = 2) + gender
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    523 491.54
## 2    520 481.67  3    9.8716 3.5524 0.01436 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When conducting an f-test, we can reject the null that the coefficients of age incorporated into the model are equal to 0 (P = .01436). Therefore, the linear model of bmi with age and gender is not sufficient (when comparing models 3 and 4) to explaining the effects of BMI on TC when controlling for age and gender.

3. Finally, tested a model that flexibly accounts for age(incorprating as potential confounder, without concern for it's actual relationship with tc)

```
mod5=gam(tc~bmi+s(age)+gender, data=homework)
summary(mod5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## tc ~ bmi + s(age) + gender
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.74960    0.25057  18.955  < 2e-16 ***
## bmi          0.03036    0.01021   2.973  0.00308 **
## gender       0.07562    0.08471   0.893  0.37243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df     F p-value
## s(age) 2.975  3.743 18.14 5.8e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.163   Deviance explained = 17.1%
## GCV = 0.93309  Scale est. = 0.92251   n = 527
```

```
anova(mod3,mod5)
```

```
## Analysis of Variance Table
##
## Model 1: tc ~ bmi + age + gender
```

```
## Model 2: tc ~ bmi + s(age) + gender
##   Res.Df    RSS     Df Sum of Sq      F   Pr(>F)
## 1 523.00 491.54
## 2 521.02 480.65 1.9752    10.896 5.9797 0.002822 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
modtest=gam(tc~s(bmi)+ bmi+s(age)+gender, data=homework)
```

```
summary(modtest)$r.sq
```

```
## [1] 0.1786691
```

```
summary(mod5)$r.sq
```

```
## [1] 0.1627437
```

Ultimately, the final model I chose that best explains the effects of BMI on Total Cholesterol while flexibly controlling for age and for age, included linear continous BMI and a GAM of age. While some information was gained from incorporating BMI in a GAM, the loss of interpretibility resulted in a decision to choose the simpler model. Considering the adjusted r-squared values for the model that included a GAM for bmi (.1787) was only slightly different than that including linear BMI (.1627), the information gained did not outdo the loss of interpretability. Initially, I had a linear model that incorporated BMI, age and gender. When conducting an f-test between the linear and the model with gam(age), we can reject the null that the coefficient of the gam(age) is 0 (p=.002822). Therefore, the linear model of bmi with age and gender is not sufficient to explaining the effects of BMI on total cholesterol when controlling for age and gender. The GAM was chosen over the spline with the intention of reproducibility and parsimony.

I recommend the following model:

$$E[tc] = \beta_0 + \beta_1(bmi) + \beta_2(GAM(age)) + \beta_3(gender)$$

(b) Take your final model and write 1-2 summary sentences that describe the overall results of your model in a form that could appear in a manuscript. Use American units of cholesterol in this summary. Be sure to include sufficient statistical detail (confidence intervals, p-values, decimal places, etc.), clarity of adjustment factors, and interpretation (units, direction of effect) in your sentence(s). Maximum of two sentences!

```
mod5$coefficients
```

```
## (Intercept)         bmi      gender    s(age).1    s(age).2    s(age).3
##  4.74960324  0.03035933  0.07562295  0.21240947  0.08495849 -0.01399419
##    s(age).4    s(age).5    s(age).6    s(age).7    s(age).8    s(age).9
##  0.11974188  0.03226482  0.10293138  0.04506069  0.52013707  0.21298287
```

```
summary(mod5)$p.table
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 4.74960324 0.25057468 18.9548409 2.499229e-61
## bmi         0.03035933 0.01021115  2.9731545 3.083986e-03
## gender      0.07562295 0.08471336  0.8926922 3.724341e-01
```

```
confint.default(mod5)
```

```
##                    2.5 %     97.5 %
## (Intercept)  4.25848589 5.24072060
## bmi          0.01034584 0.05037281
## gender      -0.09041218 0.24165808
## s(age).1    -0.22038164 0.64520057
## s(age).2    -0.54875836 0.71867535
## s(age).3    -0.25538647 0.22739809
## s(age).4    -0.25572948 0.49521323
## s(age).5    -0.13557885 0.20010848
## s(age).6    -0.19653993 0.40240269
## s(age).7    -0.07649483 0.16661621
## s(age).8    -0.62356985 1.66384400
## s(age).9    -0.24601685 0.67198259
```

side note: a 1 unit increase in BMI results in a (.03035933)38.67 = 1.17 mg/dL increase in tc. confidence interval of (0.01034584)*38.67 = and (0.05037281)38.67 = .400 and 1.948

A 1 unit increase in bmi is associated with 1.17 units (in mg/dL) increase in total cholesterol on average, with 95% CI [0.40,1.95] and P value =0.0031. The model evaluates the linear effect of BMI on total cholesterol while controlling non-parametrically for the function of age and for gender as potential confounders.