Homework 6

Amin Baabol

Note: There are no collaborators for this assignment

## Exercises

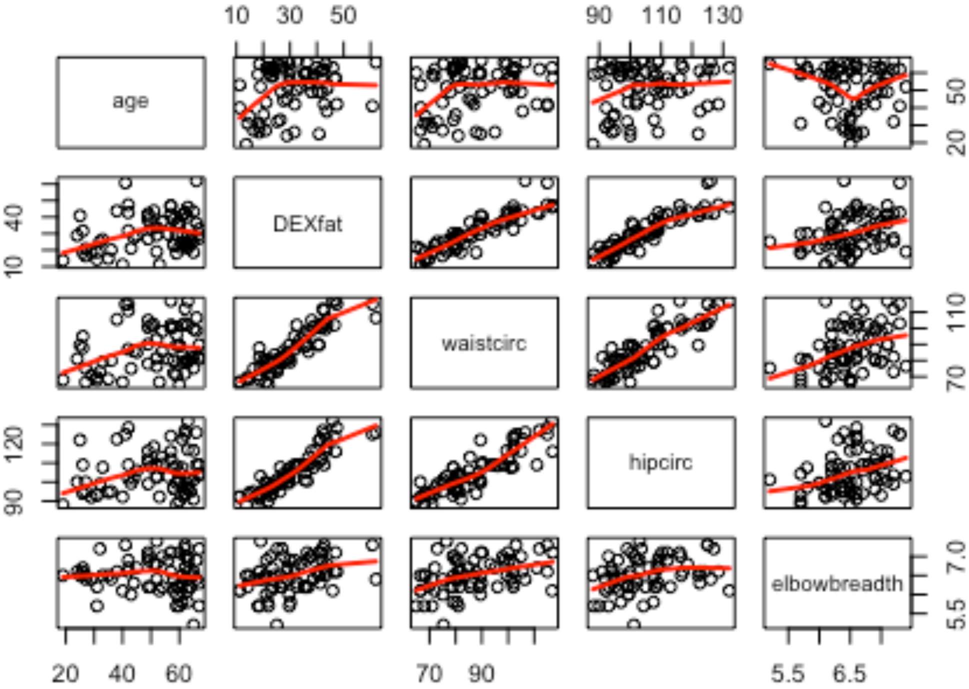
1. (Ex. 10.1 pg 207 in HSAUR, modified for clarity) Consider the **bodyfat** data from the **TH.data** package introduced in Chapter 9.

a) Use graphical methods to suggest which variables should in the model to predict body fat. (Hint: Are there correlated predictors?) Make sure to explain your reasoning.

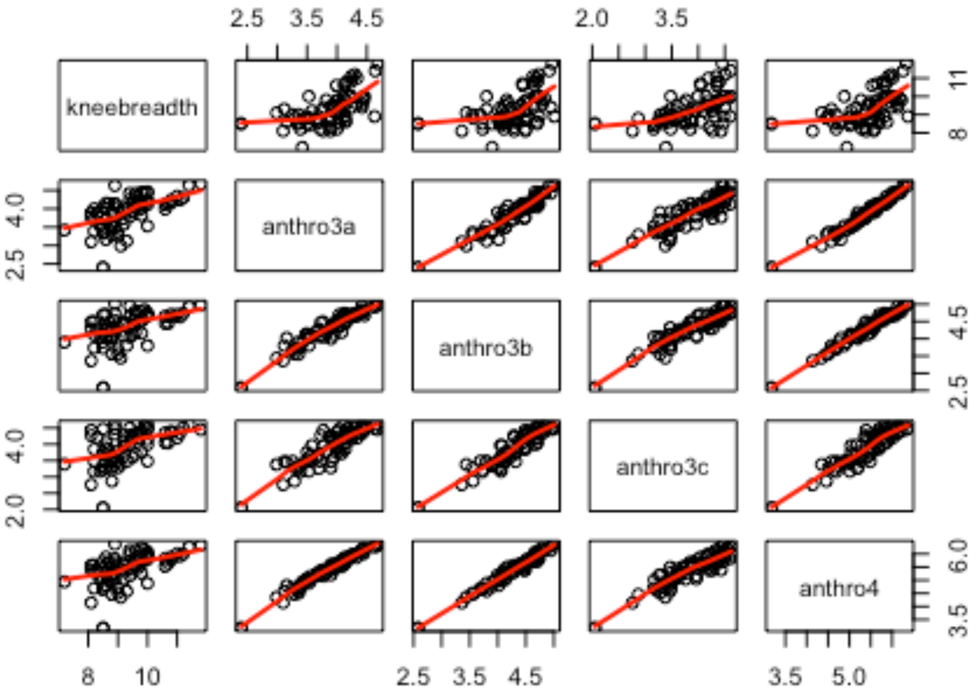
Discussion

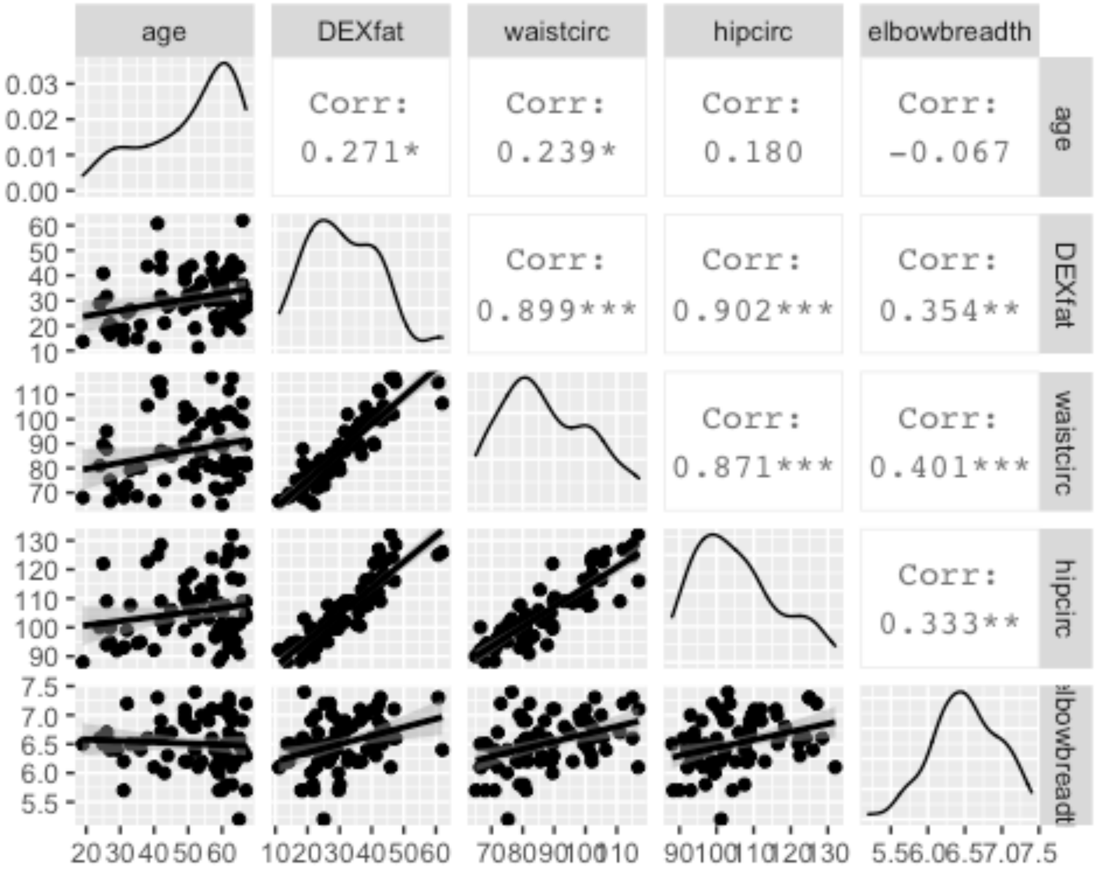
After carefully examining the plots it seems that certain predictor variables are highly correlated with each other. For example “anthro3b” and “anthro4” show strong correlation, somewhere near 0.95. This multicollinearity may present inject bias into our predictions. Therefore, I opted to remove variables that varibales that show strong correlation. While this method of variable selection through graphical means isn’t efficient it does paint a general picture of the dynamics of the variables at play.

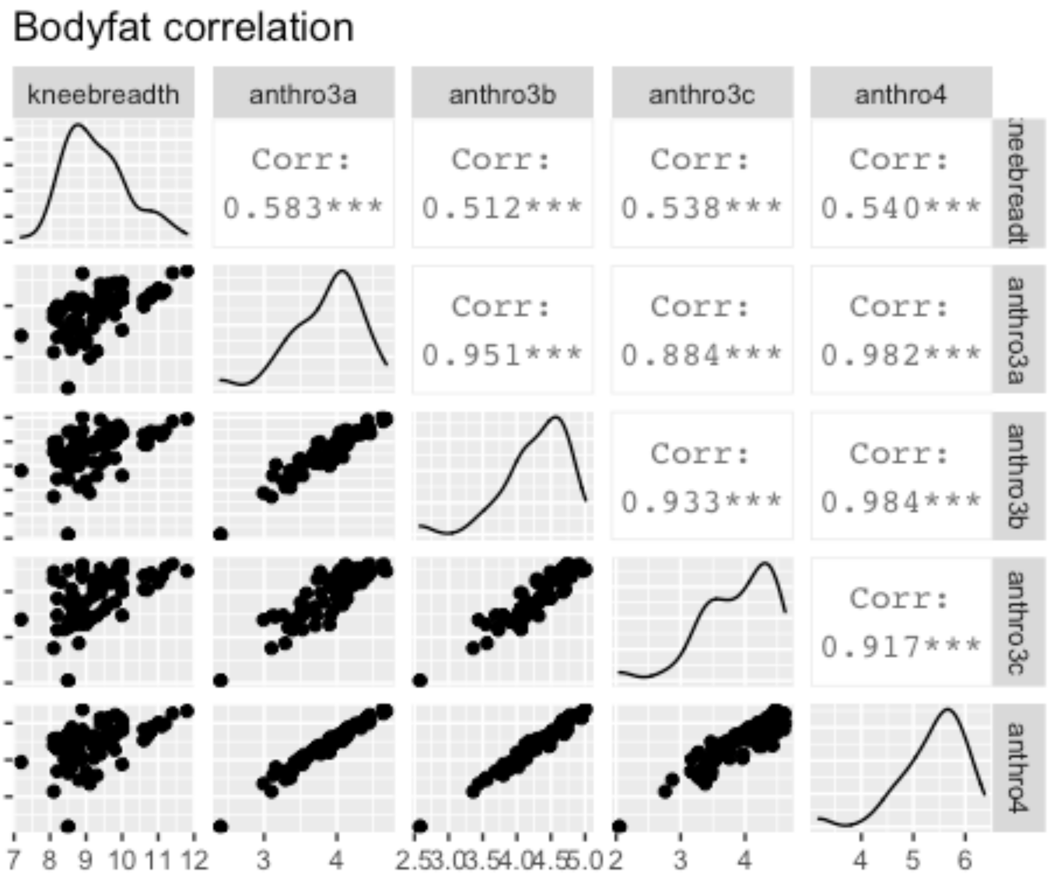
Base R: plot part 1

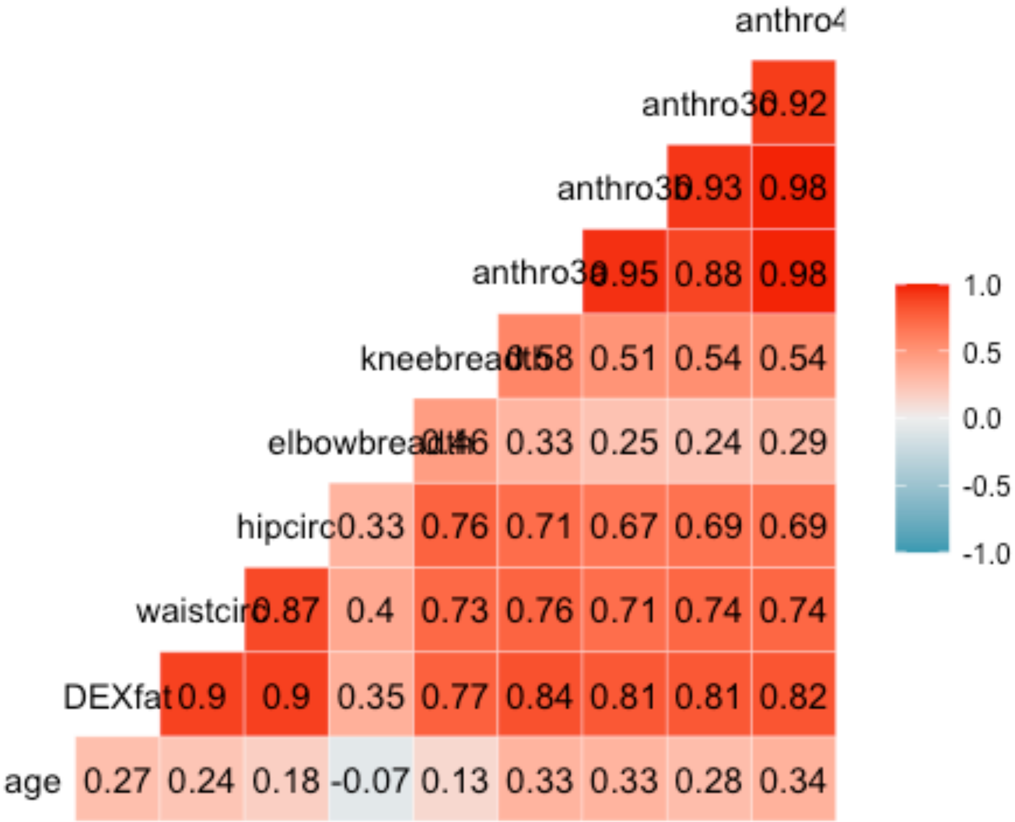


Base R: plot part 2




age
DEXfat
waistcirc
hipcirc
elbowbreadth
Corr:
0.271*
Corr:
0.239*
Corr:
0.180
Corr:
-0.067
Corr:
0.899***
Corr:
0.902***
Corr:
0.354**
Corr:
0.871***
Corr:
0.401***
Corr:
0.333**
0.03
0.02
0.01
0.00
60
50
40
30
20
10
110
100
90
80
70
130
120
110
100
90
7.5
7.0
6.5
6.0
5.5
20 30 40 50 60
10 20 30 40 50 60
70 80 90 100 110
90 100 110 120 130
5.5 6.0 6.5 7.0 7.5


Bodyfat correlation
kneebreadth
anthro3a
anthro3b
anthro3c
anthro4
Corr:
0.583***
Corr:
0.512***
Corr:
0.538***
Corr:
0.540***
Corr:
0.951***
Corr:
0.884***
Corr:
0.982***
Corr:
0.933***
Corr:
0.984***
Corr:
0.917***
neebreadt
anthro3a
anthro3b
anthro3c
anthro4
0.4
0.3
0.2
0.1
0.0
4
3
5.0
4.5
4.0
3.5
3.0
2.5
4
3
2
6
5
4
7 8 9 10 11 12
3 4
2.53.03.54.04.55.0
2 3 4
4 5 6



b) For feasability of the class, fit a generalised additive model assuming normal errors using the following code.

- Assess the **summary()** and **plot()** of the model (don't need GGPLOT for a plot of the model). Are all covariates informative? Should all covariates be smoothed or should some be included as a linear effect?
- Report GCV, AIC, and total model degrees of freedom. Discuss how certain you are that you have a reasonable summary of the actual model flexibility.
- Produce a diagnostic plot using **gam.check()** function. Are any concerns raised by the diagnostic plot?
- Write a discussion on all of the above points.

Discussion

To illustrate the effects predictor variables on DEXfat, I off started the feasibility assessment by fitting the generalized additive model only selecting to include the remaining variables from part a of this exercise. bodyfat_gam <- gam(DEXfat~ s(age) + s(waistcirc) + s(hipcirc) + s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) + s(anthro3c), data = body.fat)

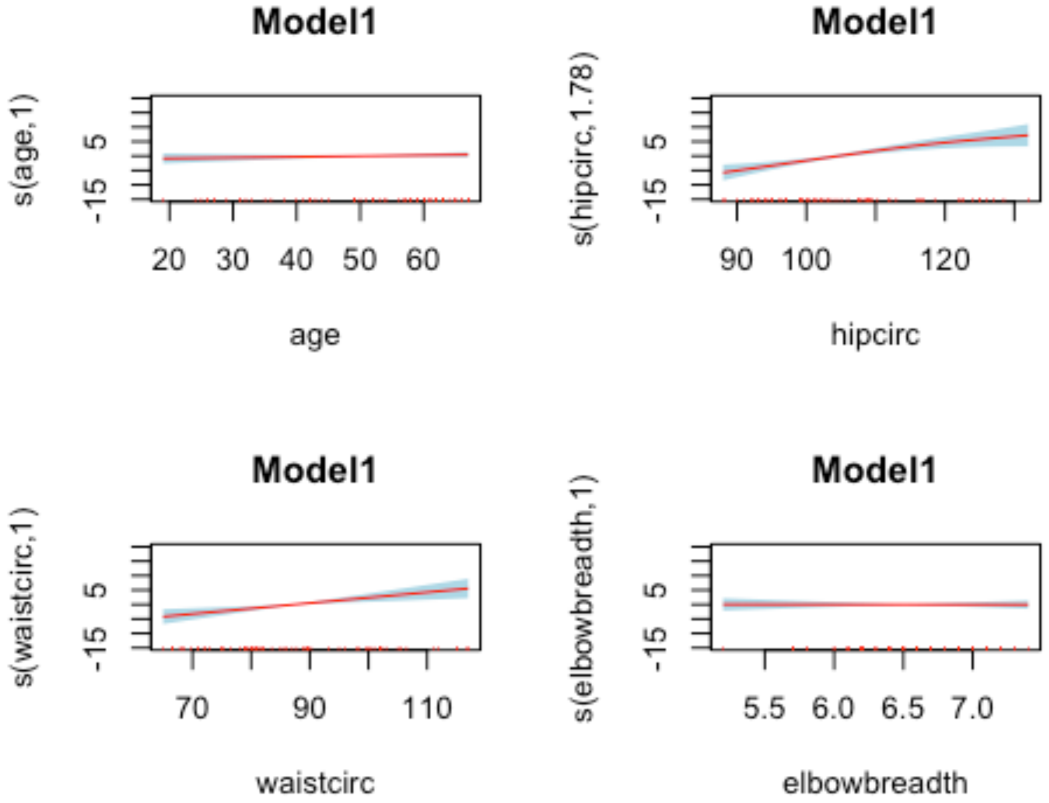
According to the statistics of the model summary only the following parametric and smooth terms have statistical significance: -the intercept -waistcirc -hipcirc -kneebreadth -anthro3a

Furthermore, the effective degree of freedom indicates the complexity of the smooth terms indicates that most terms are have EDF of 1 means most of these terms may be linear except for kneebreadth and anthro3c.it’s not enough to only check the p-values for the smooth terms during the variable selection process,visually inspecting the partial effect plots, it appears that a horizontal line might be able to fit through the partial effect plots for some of the covariates, mainly age and elbowbreadth. Fitting a horizontal line in the confidence interval means the smooth term is hardly explaining the changes in the response and therefore should NOT be included.

Moreover, the model reports: -GCV of 8.4354 -AIC of 345.708 -Adjusted R-squared of 0.953 with a deviance explained of 96.7% -Total degrees of Freedom of 21.57091

While the adjusted R-squared and the gcv look good though the AIC is relatively high and the degrees of freedom doesn’t match the number of smoothing terms, however, running the gam.check function we see a few issues arise. Firstly, although the plot converges after 41 iterations, we see that the explanatory variables -age -elbowbreadth -anthro3c have low p-values. Interestingly enough, these variables showed high p-values during the significance of the smooth terms approximation but in the model diagnostics they’re showing low p-values which means the dimensions of the basis for the smooth are too low which may potentially lead to over-smoothing, hence they should either be dropped or should NOT be smoothed. Furthermore, in the diagnostics plots, we can see that the residuals don’t follow the line well which may be an issue with our normality assumption. Moreover, the histogram shows a slight right sknewness but nothing of concern. The residuals vs fitted plot indicate a good degree of randomness while the response vs the fitted plot shows a good linear relationship. All is all, the diagnostic plots show that while the GAM model is moderately adequate it badly need further improvements.

```
## 
## Family: gaussian 
## Link function: identity 
## 
## Formula:
## DEXfat ~ s(age) + s(waistcirc) + s(hipcirc) + s(elbowbreadth) + 
##     s(kneebreadth) + s(anthro3a) + s(anthro3c)
## 
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  30.7828     0.2847   108.1   <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)          1.000  1.000  0.956 0.333043    
## s(waistcirc)    1.000  1.000 10.821 0.001885 ** 
## s(hipcirc)      1.775  2.235  9.917 0.000171 ***
## s(elbowbreadth) 1.000  1.000  0.001 0.972248
## s(kneebreadth)  8.754  8.960  6.180 8.35e-06 ***
## s(anthro3a)     1.000  1.000 12.966 0.000751 ***
## s(anthro3c)     7.042  8.041  1.798 0.100906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.953   Deviance explained = 96.7%
## GCV = 8.4354  Scale est. = 5.7538    n = 71
```

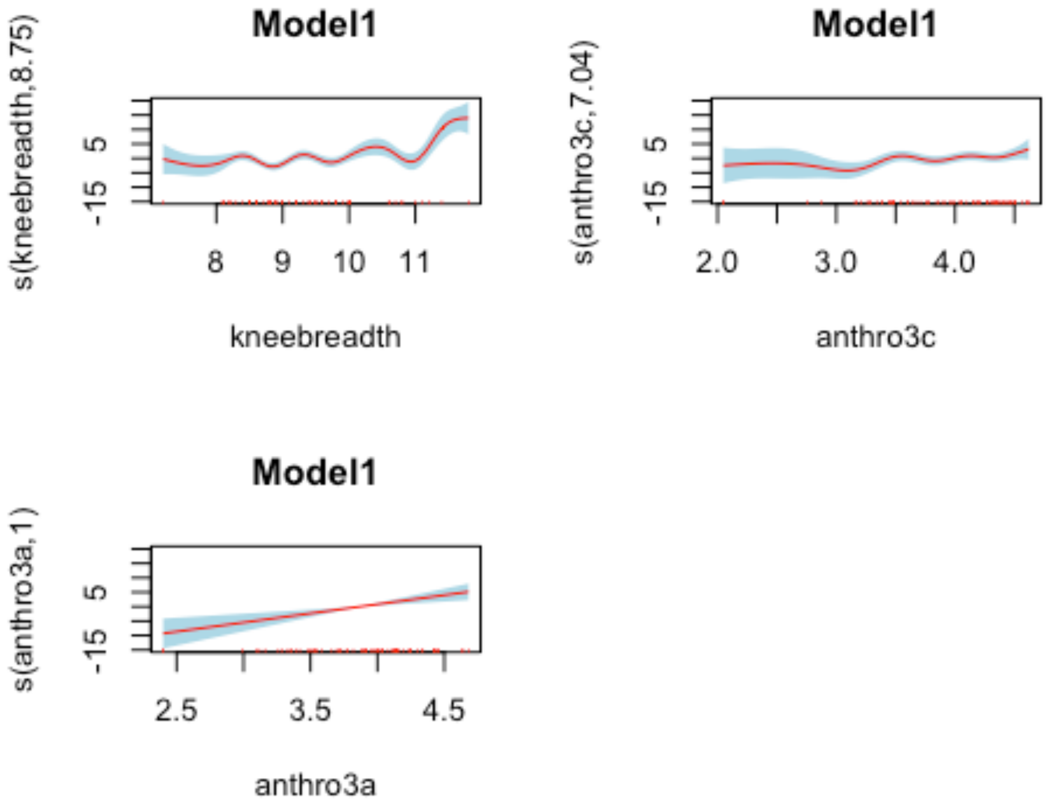


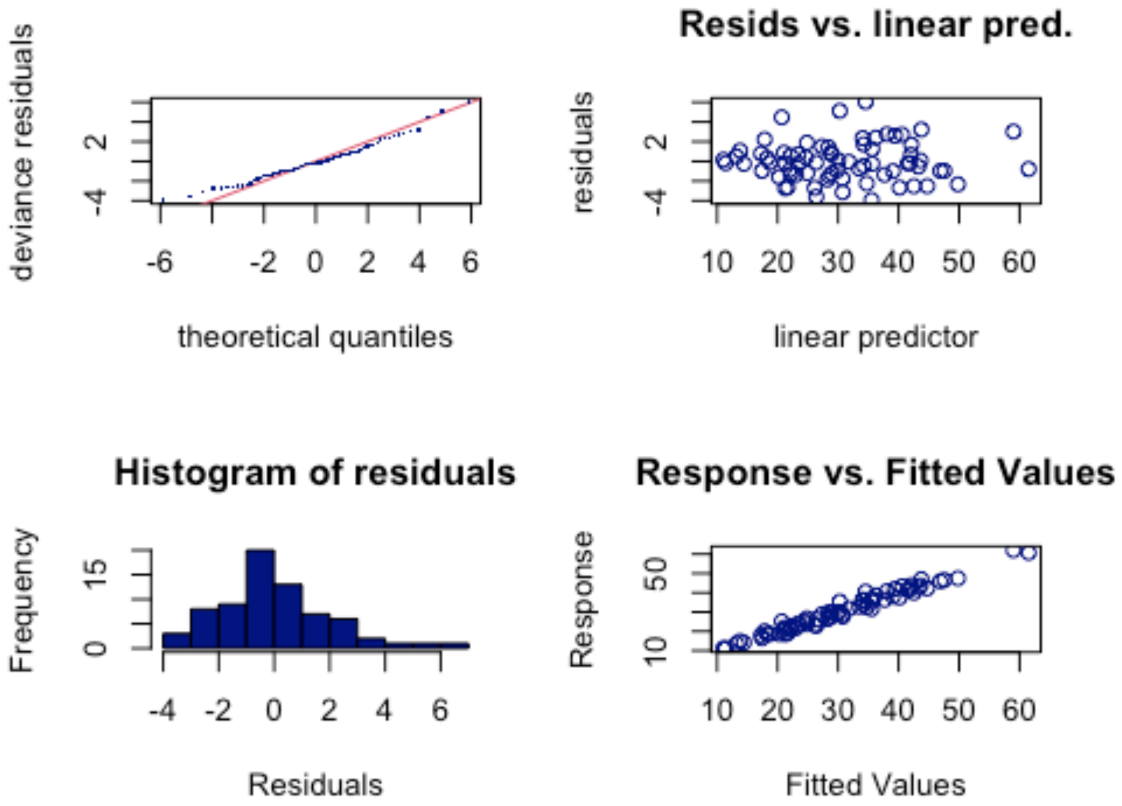
```
##   GCV.Cp
## 8.435412

## [1] 345.708

## [1] 21.57091

## [1] 0.9528156
```


Model1
s(kneebreadth,8.75)
-15
5
8
9
10
11
kneebreadth
Model1
s(anthro3c,7.04)
-15
5
2.0
3.0
4.0
anthro3c
Model1
s(anthro3a,1)
-15
5
2.5
3.5
4.5
anthro3a



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 41 iterations.
## The RMS GCV score gradient at convergence was 2.767255e-07 .
## The Hessian was positive definite.
## Model rank =  64 / 64
##
## 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(age)            9.00 1.00    0.81   0.035 *
## s(waistcirc)      9.00 1.00    0.94   0.255
## s(hipcirc)        9.00 1.78    1.02   0.510
## s(elbowbreadth) 9.00 1.00    0.81   0.045 *
## s(kneebreadth)  9.00 8.75    1.08   0.680
## s(anthro3a)       9.00 1.00    1.09   0.765
## s(anthro3c)       9.00 7.04    0.89   0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c) Fit the model below, note that some insignificant variables have been removed and some other variables are no longer smoothed. Report the summary, plot, GCV and AIC.

Discussion

Running the reduced model with few variables included we get a slightly better model than the first model we ran in part b. it’s worth noting that the following variables are smoothing terms in this model: -waistcirc -kneebreadth -anthro3c

while the following variables were used as linear effects: -Intercept -waistcirc -anthro3a

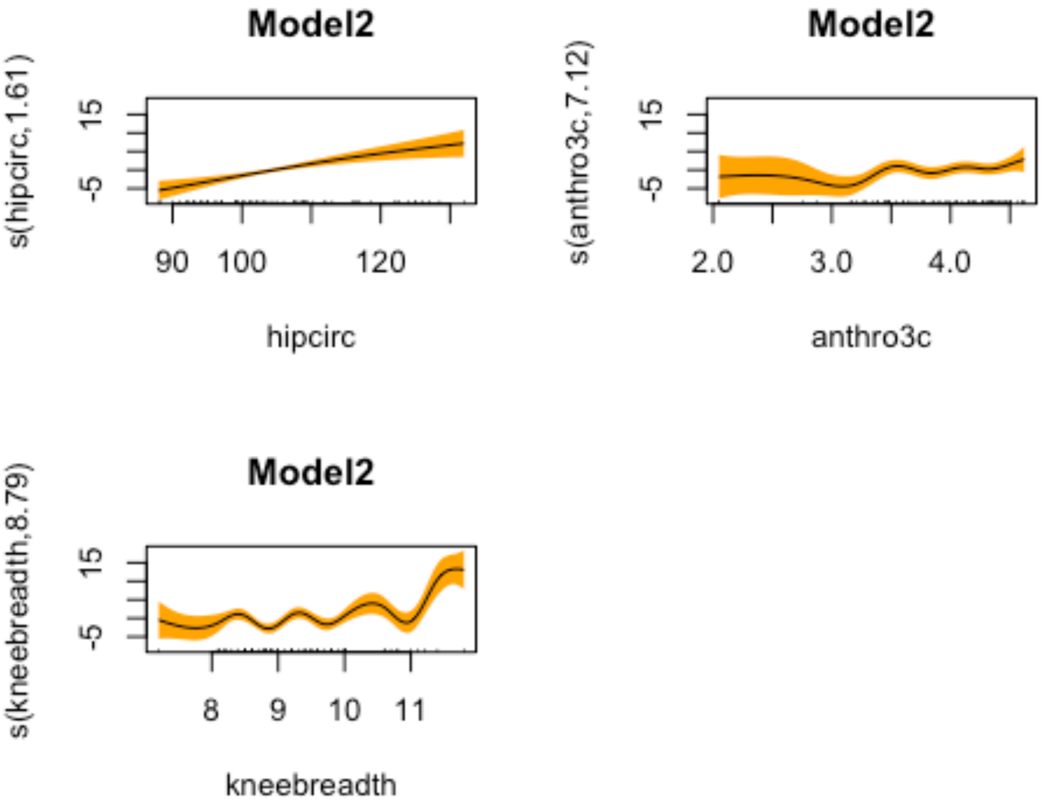
According to the p-values of the model statistics the intercept plays no statistically significant role. The plot also shows anthro3c and kneebreadth having knots corresponding to their effective degrees of freedom while hipcirc shows slightly linear line, this makes sense because hipcirc has only 1.6 effective degrees of freedom. So, I expected something in between a linear and a 2-order polynomial line. The GCV,AIC and total degrees of freedom are lower at 7.946447, 343.2562, 17.52001 respectively. This indicates an overall improvement of the second model compared to our initial model.

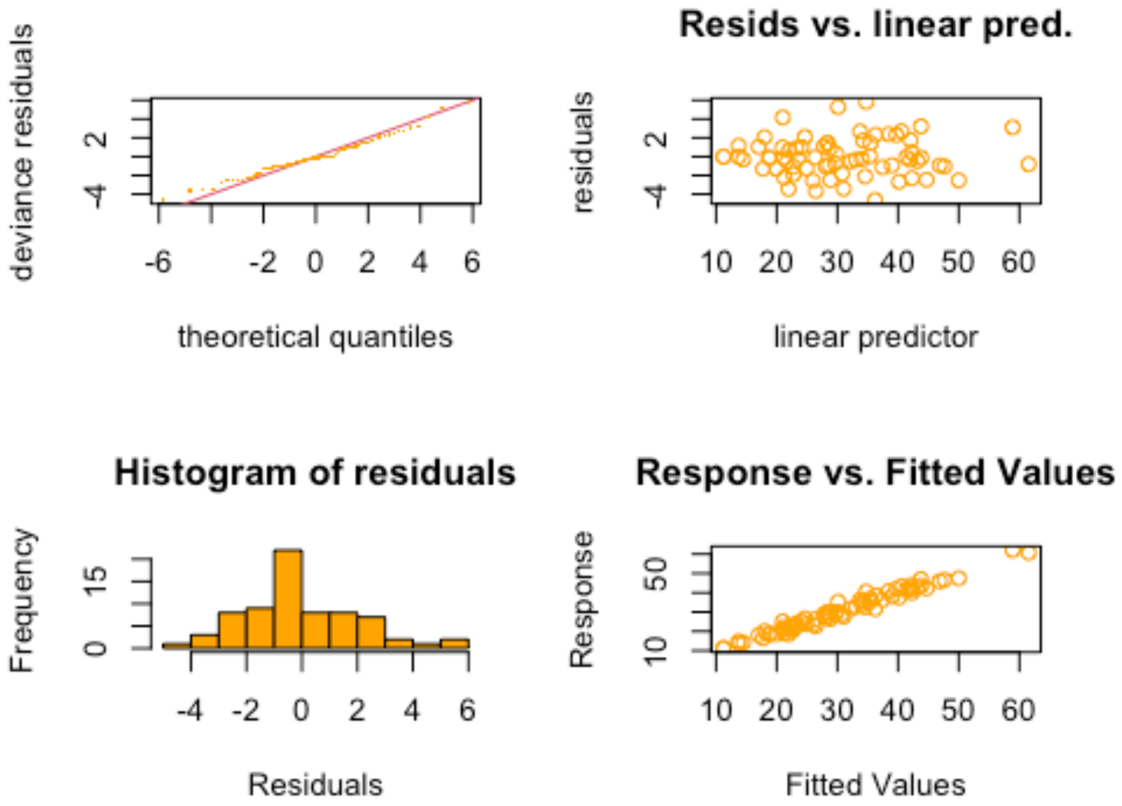
```
## 
## Family: gaussian 
## Link function: identity 
## 
## Formula:
## DEXfat ~ waistcirc + s(hipcirc) + s(kneebreadth) + anthro3a + 
##     s(anthro3c)
## 
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -13.19588    7.12570  -1.852 0.069897 .  
## waistcirc     0.19654    0.05425   3.623 0.000676 ***
## anthro3a      6.92774    1.63128   4.247 9.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##                  edf Ref.df      F  p-value    
## s(hipcirc)     1.610  2.010 10.910 0.000115 ***
## s(kneebreadth) 8.793  8.970  6.780 6.07e-06 ***
## s(anthro3c)    7.117  8.103  2.126 0.049342 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =  0.954   Deviance explained = 96.7%
## GCV = 7.9464  Scale est. = 5.6498    n = 71
```

```
##   GCV.Cp 
## 7.946447
```

```
## [1] 343.2562
```

```
## [1] 17.52001
```


Model2
s(hipcirc,1.61)
15
-5
90
100
120
hipcirc
Model2
s(anthro3c,7.12)
15
-5
2.0
3.0
4.0
anthro3c
Model2
s(kneebreadth,8.79)
15
-5
8
9
10
11
kneebreadth



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 24 iterations.
## The RMS GCV score gradient at convergence was 0.0001386163 .
## The Hessian was positive definite.
## Model rank =  30 / 30
##
## 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(hipcirc)     9.00 1.61    1.01    0.49
## s(kneebreadth) 9.00 8.79    1.06    0.66
## s(anthro3c)    9.00 7.12    0.91    0.17
```

d) Again fit an additive model to the body fat data, but this time for a log-transformed response. Compare the three models, which one is more appropriate? (Hint: use AIC, GCV, residual plots, etc. to compare models).

Discussion

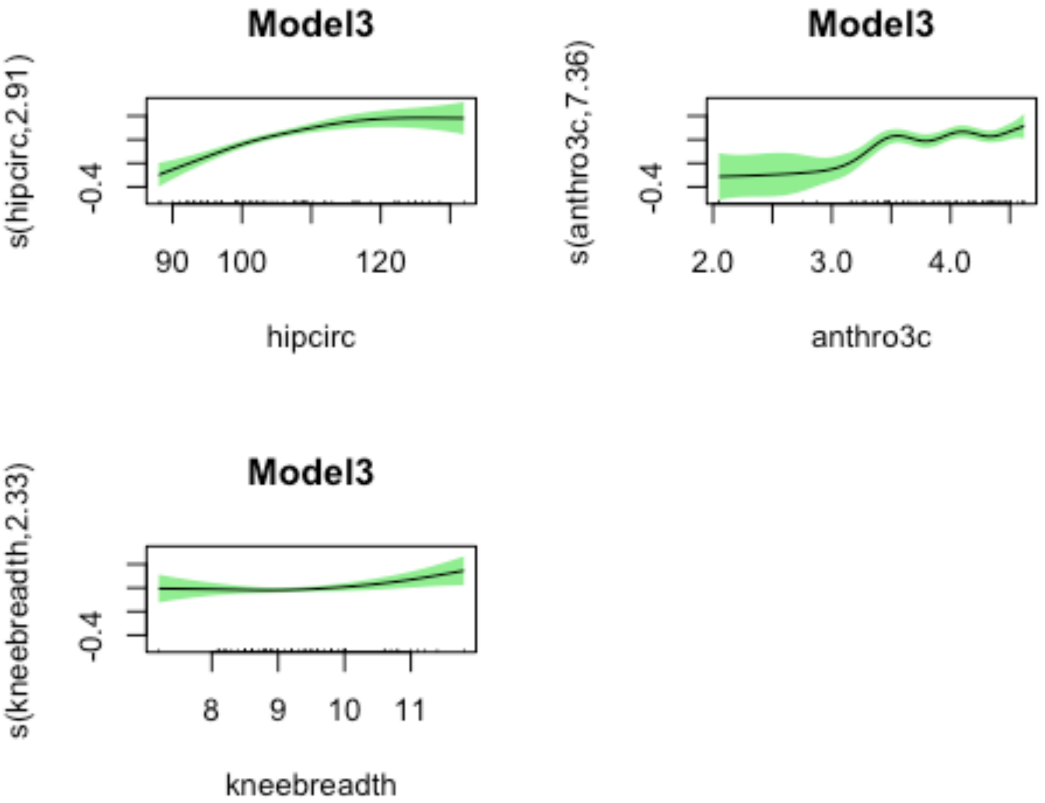
This third model uses the same explanatory variables to assess their effect on the log-transformed response variable DEXfat. I started by creating a new column for the log transformed values of DEXfat in the original dataset and appended this column into the dataset. Running the summary statistics, we quickly notice that all the parametric coefficient are statistically significant in explaining the log-transformed response variable. However, in the smooth terms we see that hipcirc and anthro3c have extremely low p-values. This might be a bit of a concern because low p-values in the smooth terms means the residuals are not randomly distributed enough, hence not enough basis functions. Interestingly enough, we also see that adjusted R squared hasn't changed noticeably at 0.952 which means it only dropped by 0.02. The GCV, AIC and total degrees of freedom are respectively at 0.0088137,-136.47 and 12.59274. Furthermore, the residuals appear to be better following the line in the qq-plot, while the histogram showing better normally distributed residuals from the right skewness present in first and second models.Overall, the residual plots as well as the other assessments generally indicate an improved and more adequate model that does a better job at explaining the log-transformed response variable compared to the first and second models.

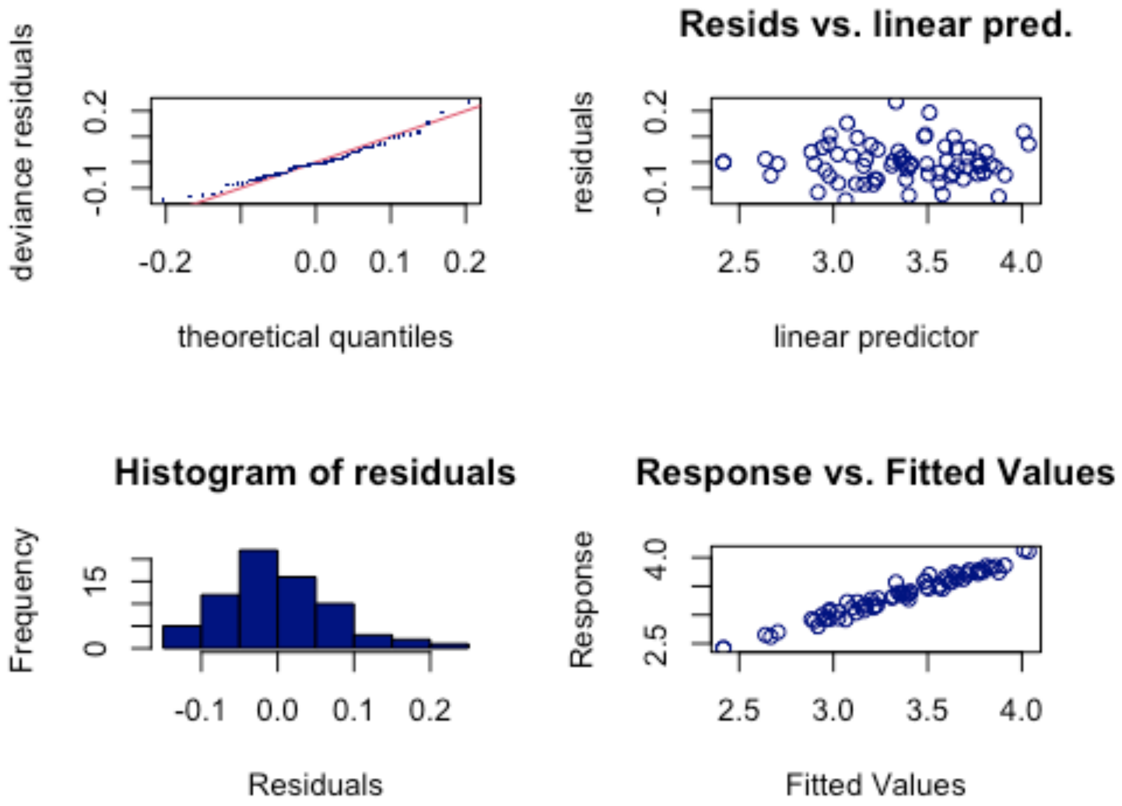
```
## 
## Family: gaussian 
## Link function: identity 
## 
## Formula:
## Log.Transformed.Response ~ waistcirc + s(hipcirc) + s(kneebreadth) + 
##     anthro3a + s(anthro3c)
## 
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.139779   0.237083   9.025  1.8e-12 ***
## waistcirc   0.004418   0.001806   2.447 0.017610 *  
## anthro3a    0.215488   0.054600   3.947 0.000226 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##                  edf Ref.df      F p-value    
## s(hipcirc)     2.909  3.616 11.828 2.1e-06 ***
## s(kneebreadth) 2.325  2.962  2.027 0.12842    
## s(anthro3c)    7.358  8.263  4.678 0.00018 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =  0.952   Deviance explained = 96.2%
## GCV = 0.0088137  Scale est. = 0.006878  n = 71

##      GCV.Cp 
## 0.008813659

## [1] -136.47

## [1] 12.59274
```


Model3
s(hipcirc,2.91)
-0.4
90
100
120
hipcirc
Model3
s(anthro3c,7.36)
-0.4
2.0
3.0
4.0
anthro3c
Model3
s(kneebreadth,2.33)
-0.4
8
9
10
11
kneebreadth



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 12 iterations.
## The RMS GCV score gradient at convergence was 9.215949e-08 .
## The Hessian was positive definite.
## Model rank =  30 / 30
##
## 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(hipcirc)      9.00 2.91    0.86   0.080 .
## s(kneebreadth) 9.00 2.33    0.83   0.075 .
## s(anthro3c)    9.00 7.36    0.99   0.430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

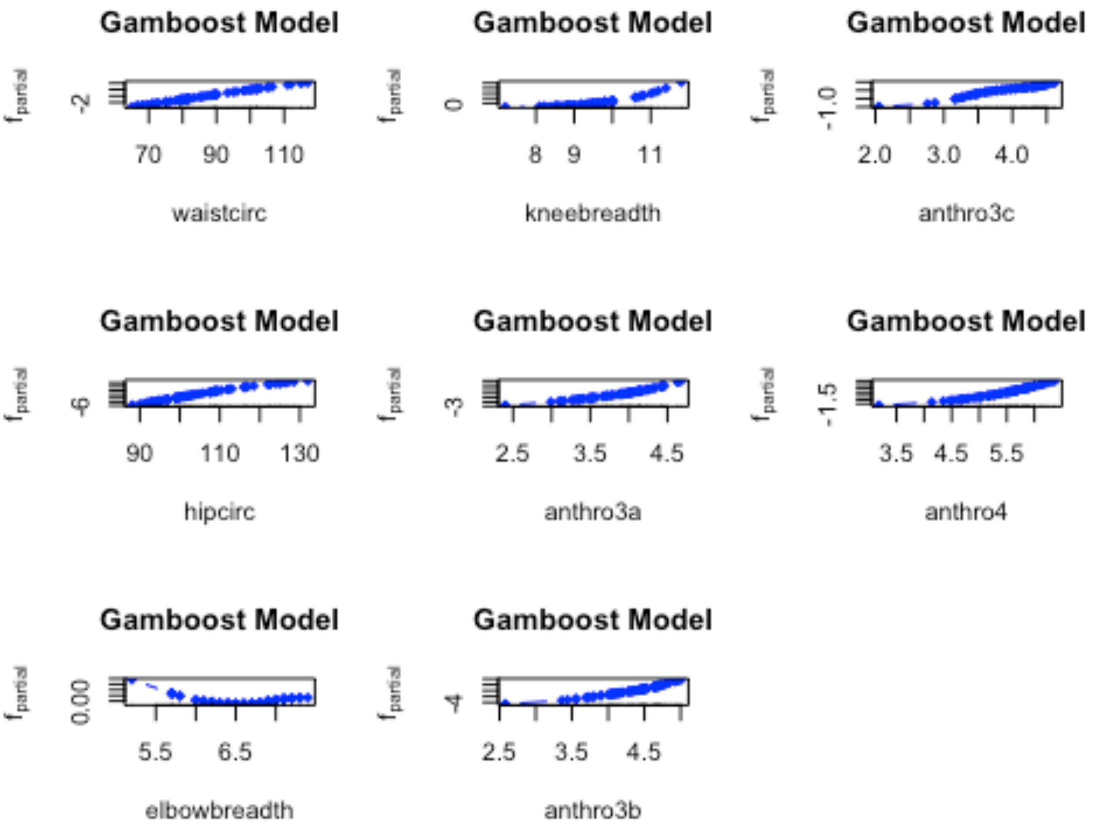
e) Run the code below to fit a generalised additive model that underwent AIC-based variable selection (fitted using the **gamboost()** function). What variable(s) was/were removed by using AIC?

Discussion

This fourth model managed to not only reduce the AIC from 345 to 3.3 but it also managed to do so without dropping too many variables in the process. In the previous models, we removed variables we thought didn't have a lot of significance in explaining the response variable. This model used AIC ranking system to drop only what is necessary and so it the explanatory variable age was dropped. We learned from our part b of the assessment that age wasn't significant but we also dropped two other variables in the process. Ultimately, our third model shows lower AIC than this fourth model. But, it's interesting to note how much smoother and "parametric-like" for lack of a better term the plots have become.

```
##
##   Model-based Boosting
##
## Call:
## gamboost(formula = DEXfat ~ ., data = bodyfat)
##
##
##   Squared Error (Regression)
##
## Loss function: (y - f)^2
##
##
## Number of boosting iterations: mstop = 51
## Step size:  0.1
## Offset:  30.78282
## Number of baselearners:  9
##
## Selection frequencies:
##  bbs(kneebreadth, df = dfbase)     bbs(anthro3b, df = dfbase)
##                     0.35294118                     0.17647059
##      bbs(hipcirc, df = dfbase)     bbs(anthro3a, df = dfbase)
##                     0.13725490                     0.11764706
##     bbs(anthro3c, df = dfbase)    bbs(waistcirc, df = dfbase)
##                     0.09803922                     0.07843137
## bbs(elbowbreadth, df = dfbase)      bbs(anthro4, df = dfbase)
##                     0.01960784                     0.01960784
```

```
## [1] 3.268173
## Optimal number of boosting iterations: 51
## Degrees of freedom (for mstop = 51): 7.637287
```



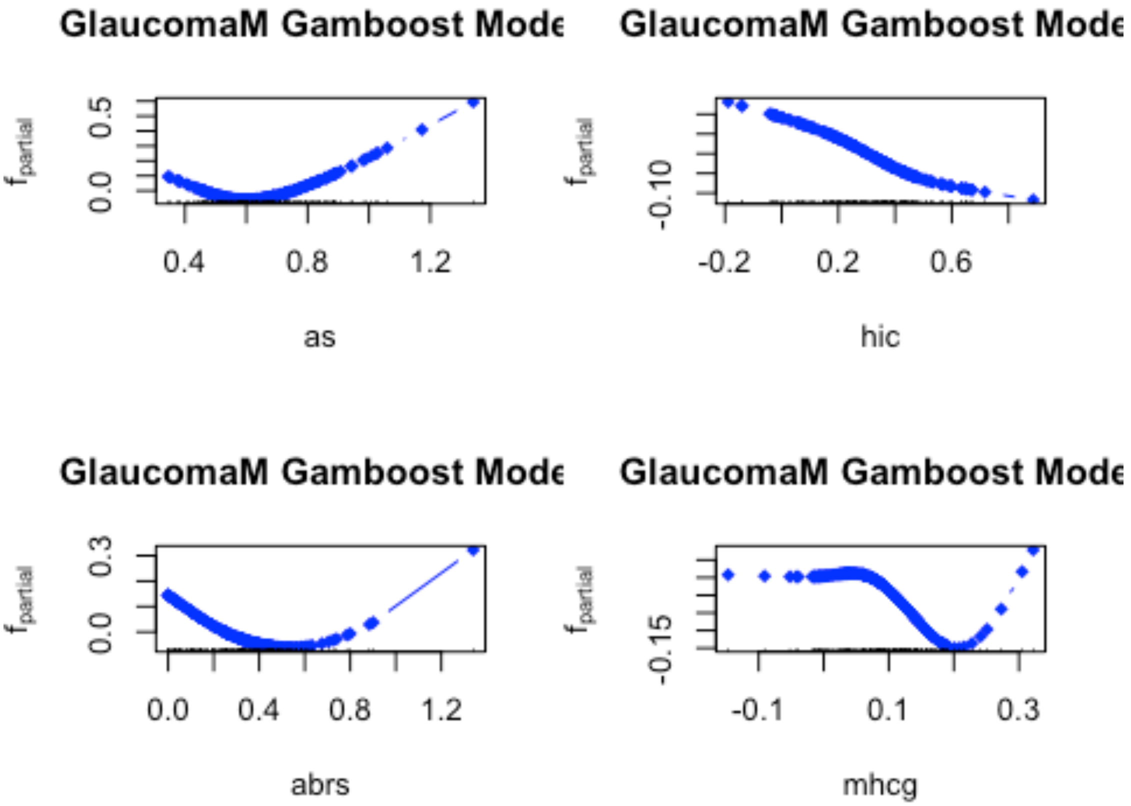
2. (Ex. 10.3 pg 208 in HSAUR, modified for clarity) Fit an additive model to the **glaucomaM** data from the **TH.data** library with *Class* as the response variable. Read the description of the dataset and the goals of the experiment. Which covariates should be in the model and what is their influence on the probability of suffering from glaucoma? (Hint: Since there are many covariates, use **gamboost()** to fit the GAM.) Make sure to provide a written summary of the model you chose and your corresponding analysis.

Discussion

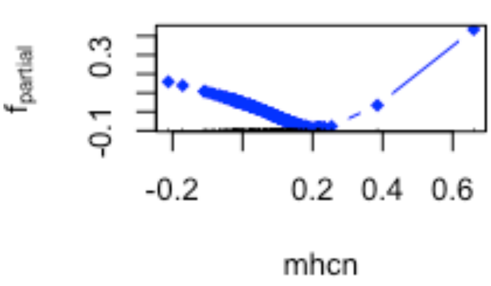
After constructing a gamboost model and running its summary, it turns out the model selected only 18 out of 63 possible explanatory variables as statistically significant. Out of the selected variables: -tmi -mhcg -vars -mhci have the highest probability in predicting the onset of glaucoma. The partial effect plots look exceptionally smooth, so much that phcg and phci appear to be linear-effects.

```
##
##   Model-based Boosting
##
## Call:
## gamboost(formula = Class ~ ., data = GlaucomaM, family = Binomial())
##
```

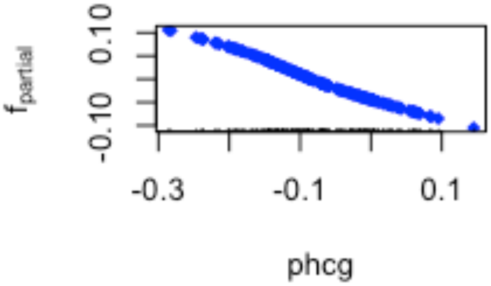
```
##
##   Negative Binomial Likelihood (logit link)
##
## Loss function: {
##      f <- pmin(abs(f), 36) * sign(f)
##      p <- exp(f)/(exp(f) + exp(-f))
##      y <- (y + 1)/2
##      -y * log(p) - (1 - y) * log(1 - p)
##  }
##
##
## Number of boosting iterations: mstop = 100
## Step size:  0.1
## Offset:  0
## Number of baselearners:  62
##
## Selection frequencies:
##  bbs(tmi, df = dfbase) bbs(mhcg, df = dfbase) bbs(vars, df = dfbase)
##                   0.17                   0.11                   0.11
## bbs(mhci, df = dfbase)  bbs(hvc, df = dfbase) bbs(vass, df = dfbase)
##                   0.10                   0.08                   0.08
##   bbs(as, df = dfbase) bbs(vari, df = dfbase)   bbs(mv, df = dfbase)
##                   0.07                   0.06                   0.04
## bbs(abrs, df = dfbase) bbs(mhcn, df = dfbase) bbs(phcn, df = dfbase)
##                   0.03                   0.03                   0.03
##  bbs(mdn, df = dfbase) bbs(phci, df = dfbase)  bbs(hic, df = dfbase)
##                   0.03                   0.02                   0.01
## bbs(phcg, df = dfbase)  bbs(mdi, df = dfbase)  bbs(tms, df = dfbase)
##                   0.01                   0.01                   0.01
```


GlaucomaM Gamboost Mode
$f_{partial}$
0.0
0.5
0.4
0.8
1.2
as
GlaucomaM Gamboost Mode
$f_{partial}$
-0.10
-0.2
0.2
0.6
hic
GlaucomaM Gamboost Mode
$f_{partial}$
0.0
0.3
0.0
0.4
0.8
1.2
abrs
GlaucomaM Gamboost Mode
$f_{partial}$
-0.15
-0.1
0.1
0.3
mhcg

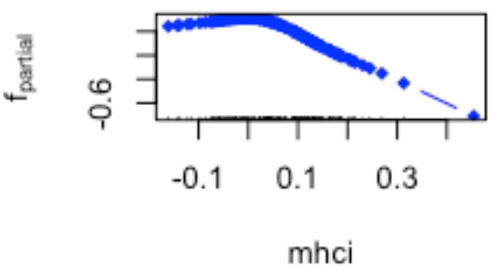
GlaucomaM Gamboost Mode



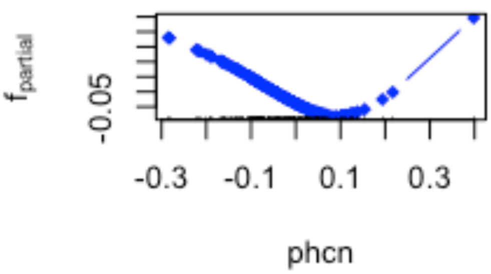
GlaucomaM Gamboost Mode

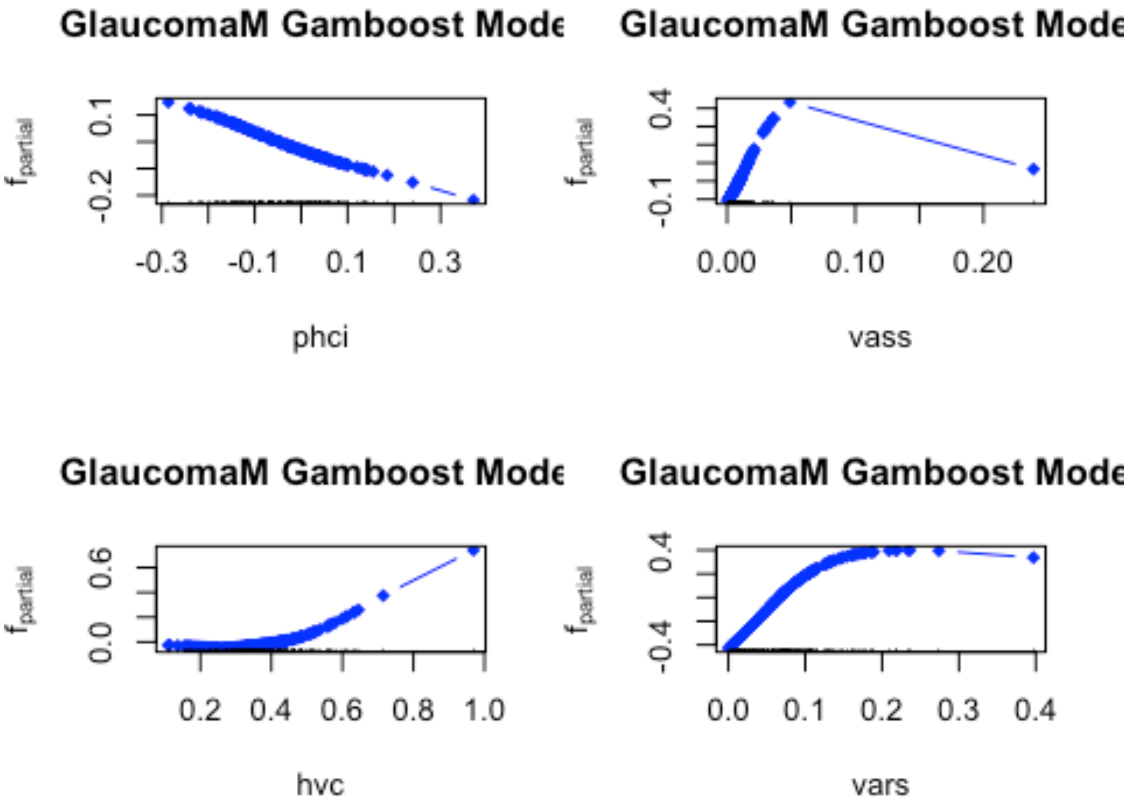


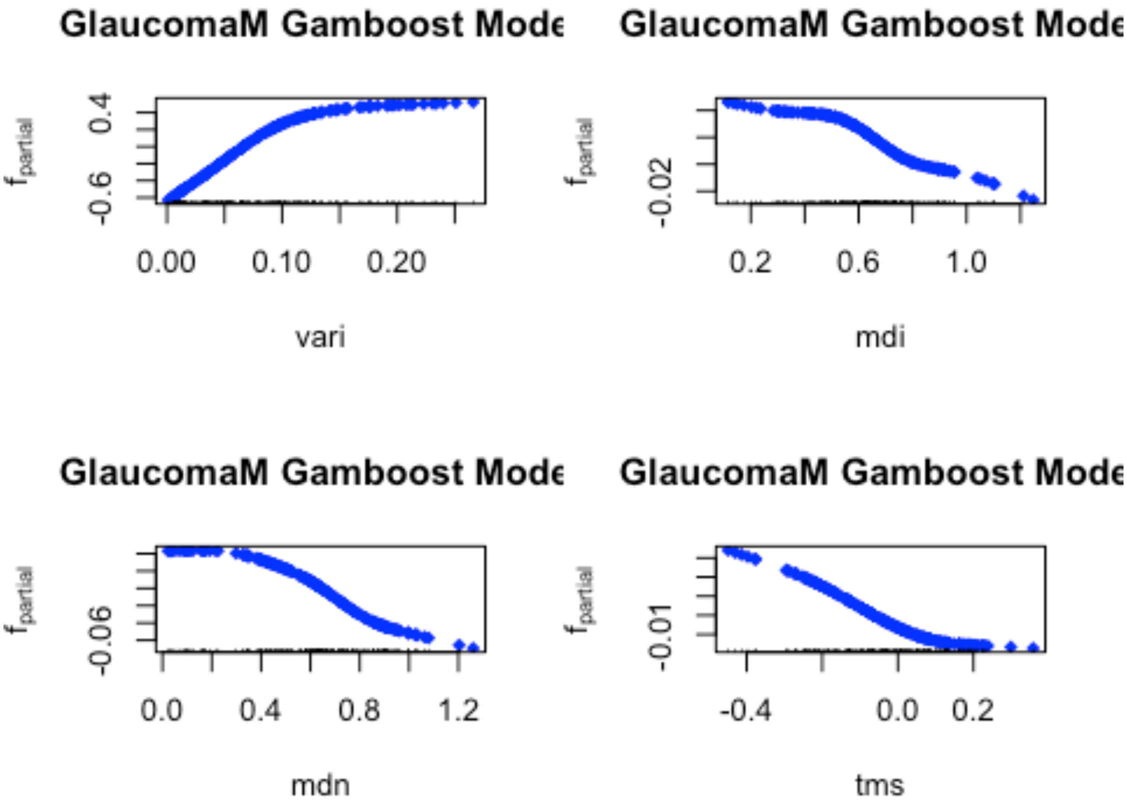
GlaucomaM Gamboost Mode

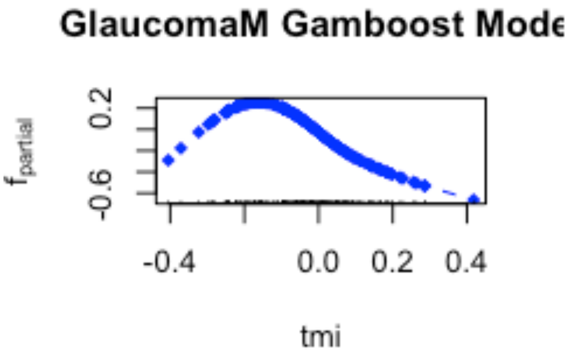


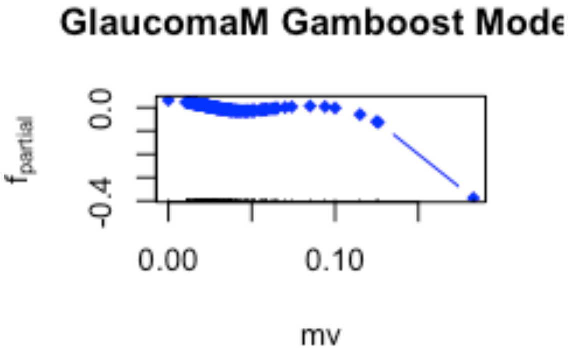
GlaucomaM Gamboost Mode




GlaucomaM Gamboost Mode
$f_{partial}$
0.1
-0.2
-0.3
-0.1
0.1
0.3
phci
GlaucomaM Gamboost Mode
$f_{partial}$
0.4
-0.1
0.00
0.10
0.20
vass
GlaucomaM Gamboost Mode
$f_{partial}$
0.6
0.0
0.2
0.4
0.6
0.8
1.0
hvc
GlaucomaM Gamboost Mode
$f_{partial}$
0.4
-0.4
0.0
0.1
0.2
0.3
0.4
vars


GlaucomaM Gamboost Mode
f_partial
0.4
-0.6
0.00
0.10
0.20
vari
GlaucomaM Gamboost Mode
f_partial
-0.02
0.2
0.6
1.0
mdi
GlaucomaM Gamboost Mode
f_partial
-0.06
0.0
0.4
0.8
1.2
mdn
GlaucomaM Gamboost Mode
f_partial
-0.01
-0.4
0.0
0.2
tms





###Works Cited

1. Michael, Semhar, and Christopher P. Saunders. “Scatterplot Smoothers and GAM” Chapter 10. 18 Oct. 2020, South Dakota State University, South Dakota State University.

2. Everitt, Brian, and Torsten Hothorn. A Handbook of Statistical Analyses Using n SECOND EDITION. Taylor and Francis Group LLC, 2010.

3. Jackson, Simon. “Visualising Residuals • BlogR.” BlogR on Svbtle, drsimonj.svbtle.com/visualising-residuals.

4. Lowhorn, J. (n.d.). Retrieved October 21, 2020, from https://rstudio-pubs-static.s3.amazonaws.com/326465_9748350bbfca41afb753211eff074761.html