# Generalized Additive Models

Mohamed Ahmed

## Exercises

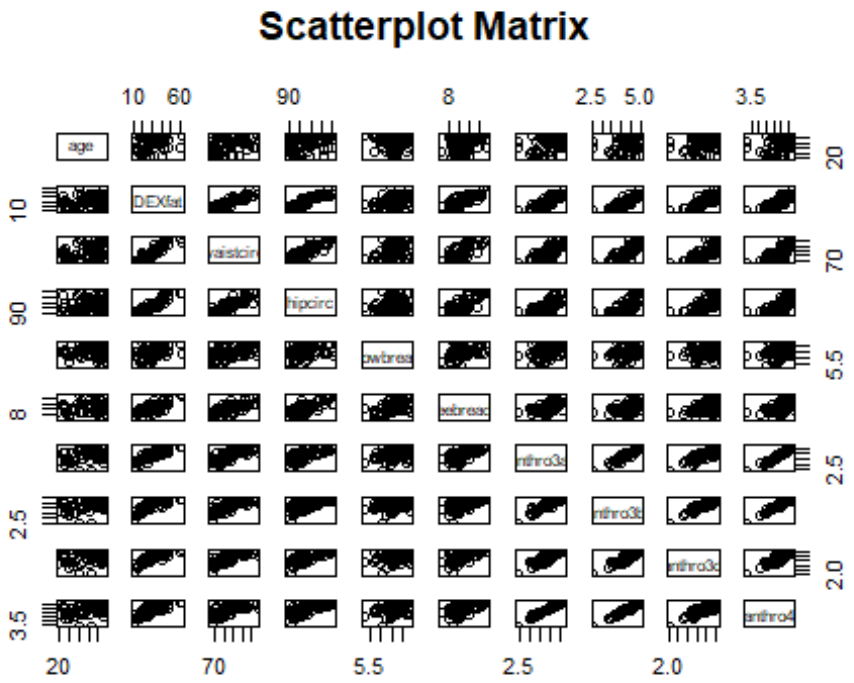
1. (Ex. 10.1 pg 207 in HSAUR, modified for clarity) Consider the **bodyfat** data from the **TH.data** package introduced in Chapter 9.

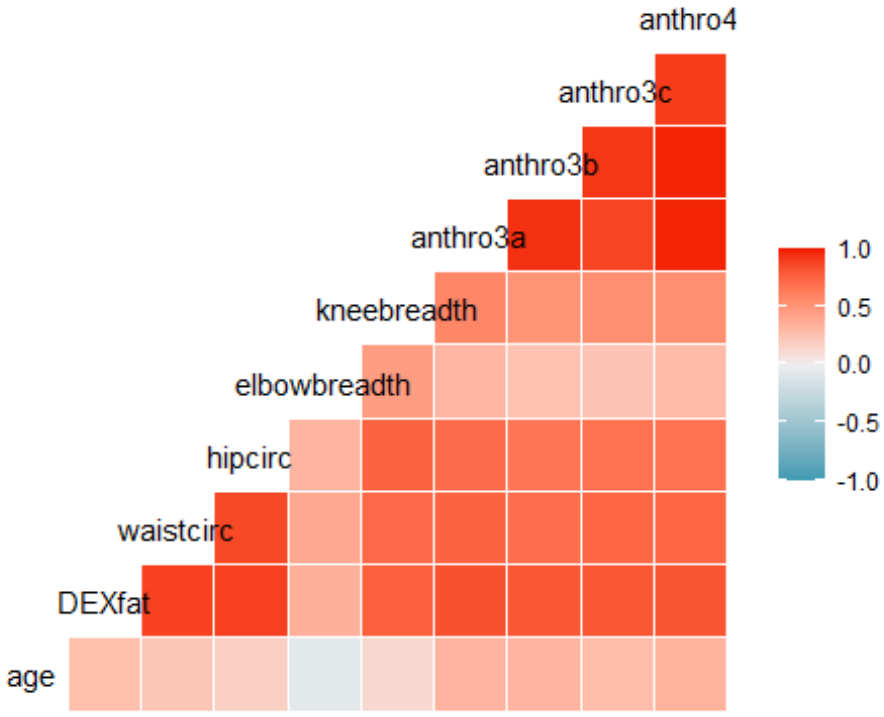
    a) Use graphical methods to suggest which variables should in the model to predict body fat. (Hint: Are there correlated predictors?) Make sure to explain your reasoning.

Answer:

We plotted the correlation matrix to examine the correlation between the explanatory variables and themselves and the correlation between the explantory variables and the response variables. We can see from the plot that there is a correlation between multiple explantory variables and the response variable DEXfat such as waistcirc and hipcirc. And we can see that there is variables that have very low correlation to DEXfat such as elbowbreadth and age. We created a table that contains the relationship between predictor variables.

However we want to avoid predictor variables that are highly correlated so,we dropped variables that have correlation that is above 0.94 and we summarized our selected predictors along with their correlations with DEXfat in the table below.




```
##                  age waistcirc hipcirc elbowbreadth kneebreadth anthro3a
anthro3b
## age          0.000      0.24    0.18        0.067        0.13     0.33
0.33
```

```
## waistcirc    0.239      0.00    0.87        0.401        0.73     0.76
0.71
## hipcirc      0.180      0.87    0.00        0.333        0.76     0.71
0.67
## elbowbreadth 0.067      0.40    0.33        0.000        0.46     0.33
0.25
## kneebreadth  0.128      0.73    0.76        0.463        0.00     0.58
0.51
## anthro3a     0.334      0.76    0.71        0.325        0.58     0.00
0.95
## anthro3b     0.332      0.71    0.67        0.253        0.51     0.95
0.00
## anthro3c     0.281      0.74    0.69        0.241        0.54     0.88
0.93
## anthro4      0.345      0.74    0.69        0.294        0.54     0.98
0.98
##              anthro3c anthro4
## age              0.28    0.34
## waistcirc        0.74    0.74
## hipcirc          0.69    0.69
## elbowbreadth     0.24    0.29
## kneebreadth      0.54    0.54
## anthro3a         0.88    0.98
## anthro3b         0.93    0.98
## anthro3c         0.00    0.92
## anthro4          0.92    0.00

## # A tibble: 6 x 2
##   `Selected Variables` Correlation
##   <chr>                       <dbl>
## 1 age                         0.271
## 2 elbowbreadth                0.354
## 3 kneebreadth                 0.768
## 4 anthro3c                    0.810
## 5 waistcirc                   0.899
## 6 hipcirc                     0.902
```

We can see from the plots, That variables

```
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.
```

Answer:

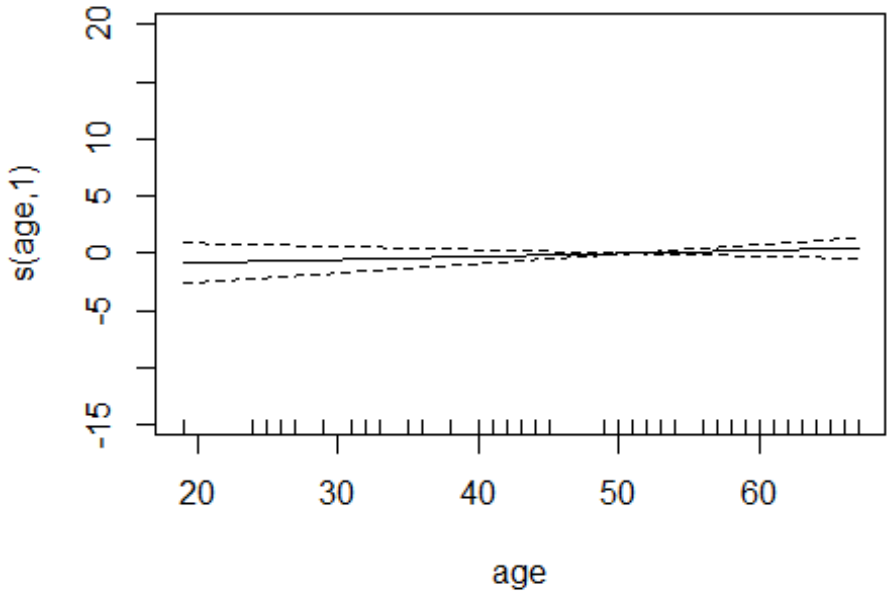
The model indicates that all the predictor variables are significant at 0.5 significance level except for **age,elbowbreadth, and anthro3c**. from the model summary and the plots, we can see thatthe variables (waistcirc, elbowbreadth, anthro3a, age) look linear and have a degrees of freedom of one. The variables **hipcirc(EDF=1.775), kneebreadth(EDF=8.754), anthro3c(EDF=7.042)** have DF above one. This indicates that these variables need smoothing of different orders based on their degrees of freedom.

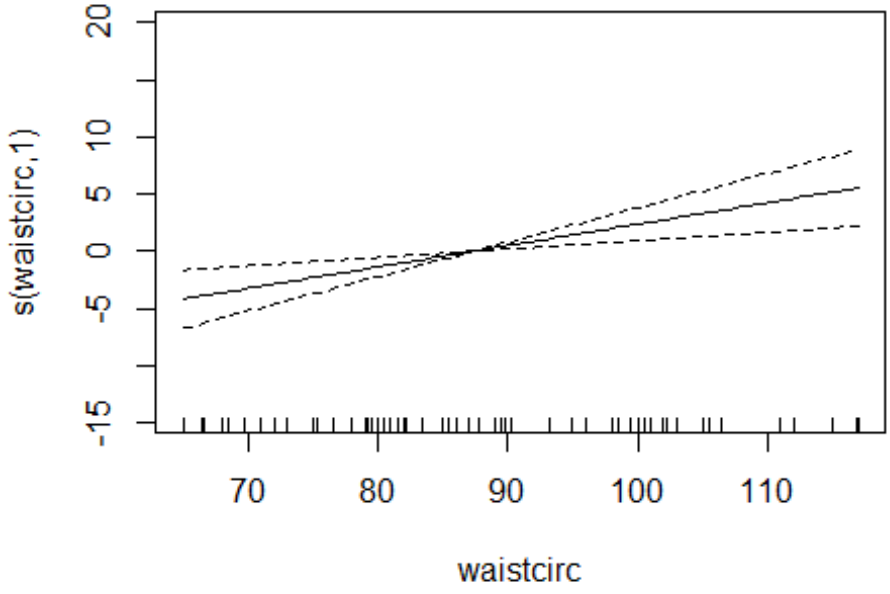
The model does not have a high GGV **GGV = 8.435412** however the model has a high **AIC = 345.708** and a high **R^2 = 0.9528156**. This indicates that the model need some needs smoothing and parameter tuning. The high R^2 means that the model explains 95% of the variation in response variables.

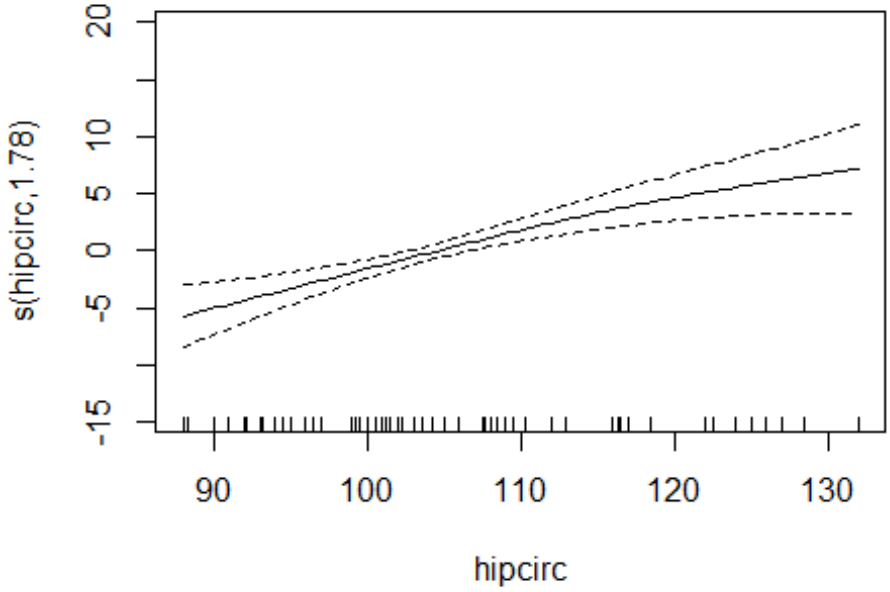
The residuals plot show that error is distributed randomly around zero. Also, the histogram show a normal distribution and the response vs fitted plot shows a linear relationship. All these plots indicate that the model is a good predictor.

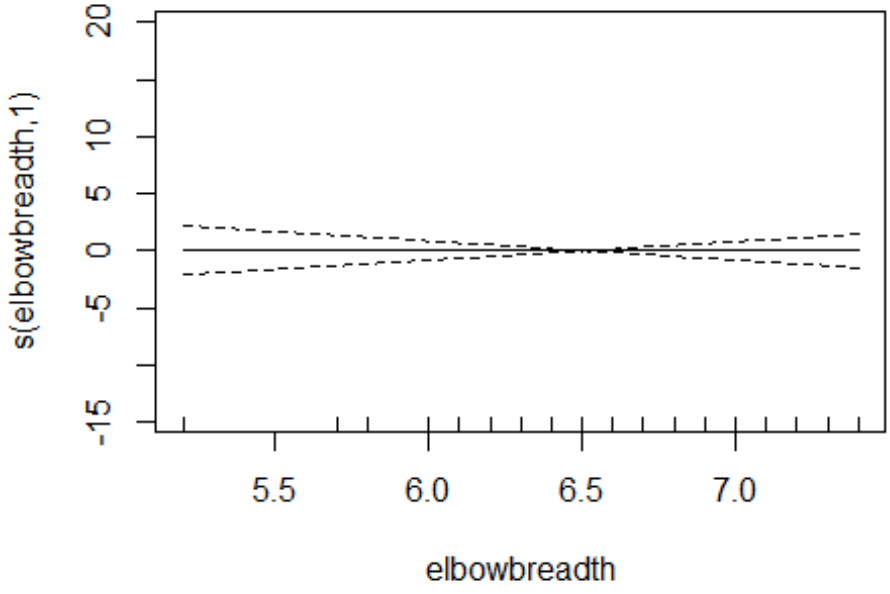
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## DEXfat ~ s(age) + s(waistcirc) + s(hipcirc) + s(elbowbreadth) +
##     s(kneebreadth) + s(anthro3a) + s(anthro3c)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.7828     0.2847   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
##
```

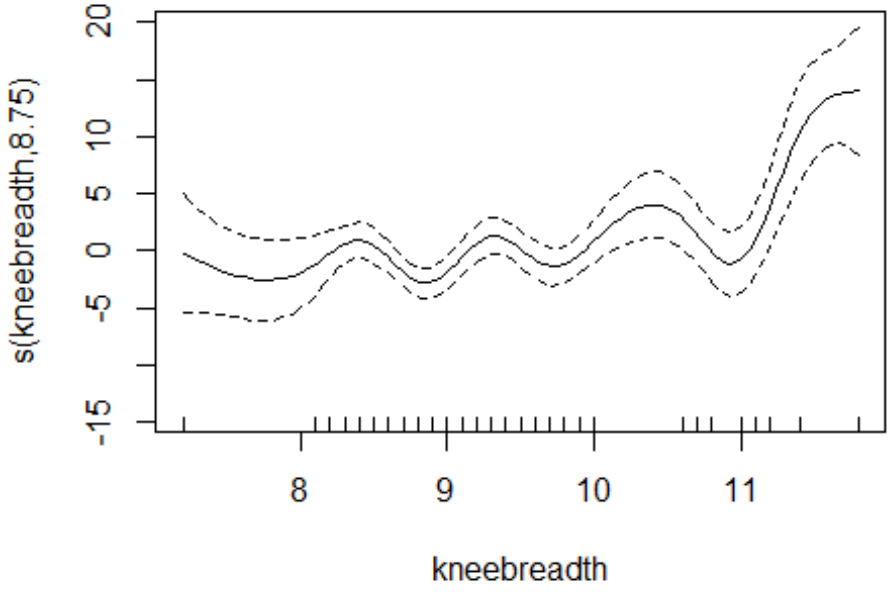
```
## R-sq.(adj) =  0.953   Deviance explained = 96.7%
## GCV = 8.4354  Scale est. = 5.7538    n = 71
```

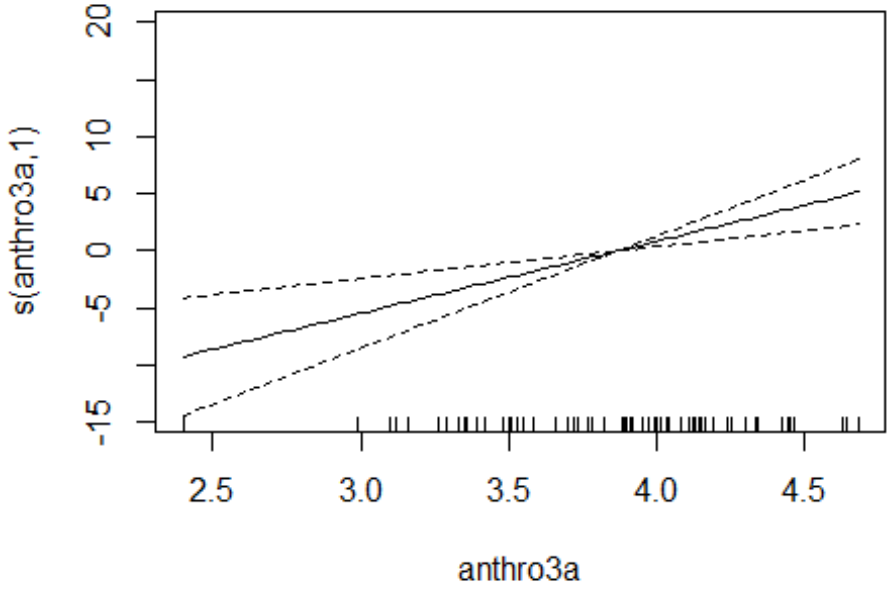


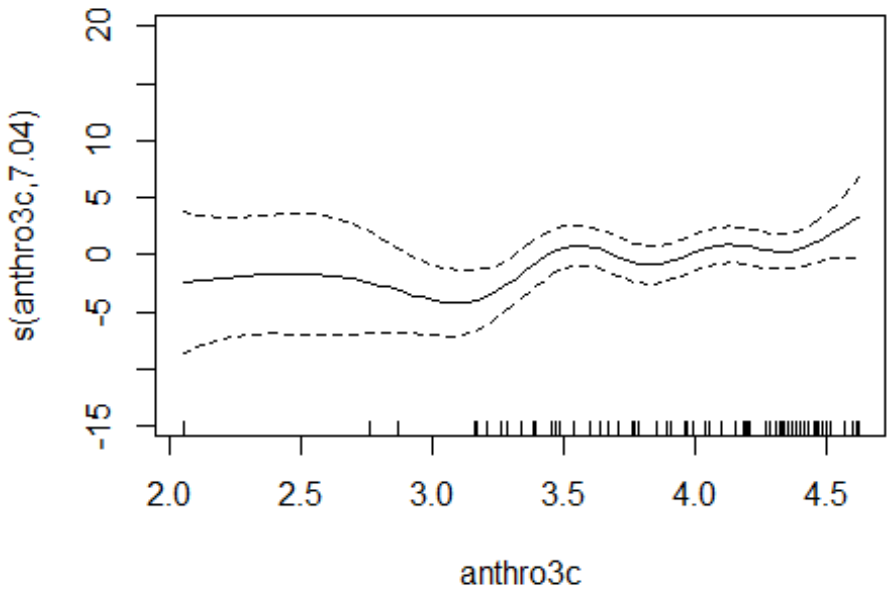




s(hipcirc,1.78)
20
10
5
0
-5
-15
90
100
110
120
130
hipcirc

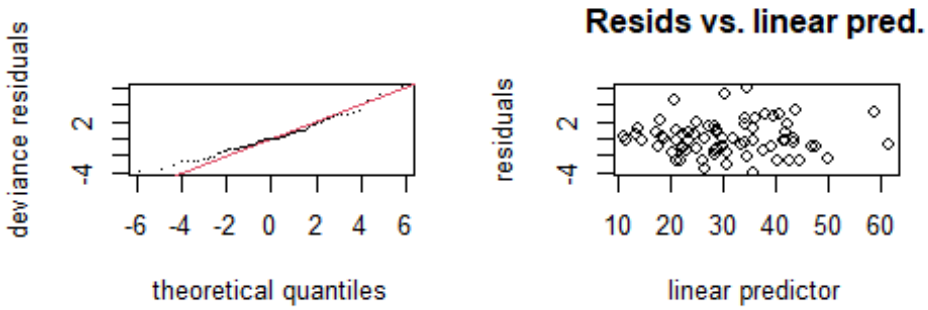

s(elbowbreadth,1)
20
10
5
0
-5
-15
5.5
6.0
6.5
7.0
elbowbreadth

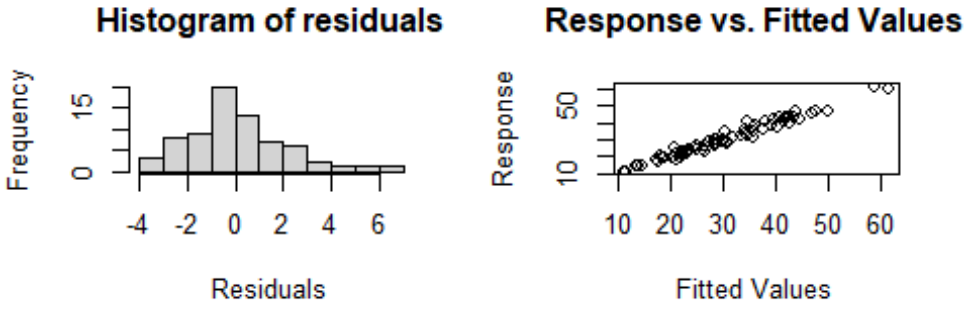

s(kneebreadth,8.75)
20
10
5
0
-5
-15
8
9
10
11
kneebreadth


s(anthro3a,1)
20
10
5
0
-5
-15
2.5
3.0
3.5
4.0
4.5
anthro3a




```
## GGV
## [1] 8.435412
## AIC
## [1] 345.708
## Adjusted R^2
## [1] 0.9528156
## DF
## [1] 21.57091
```




```
##
## 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   0.020 *
## s(waistcirc)    9.00 1.00    0.94   0.260
## s(hipcirc)      9.00 1.78    1.02   0.545
## s(elbowbreadth) 9.00 1.00    0.81   0.045 *
## s(kneebreadth)  9.00 8.75    1.08   0.685
## s(anthro3a)     9.00 1.00    1.09   0.700
## s(anthro3c)     9.00 7.04    0.89   0.110
## ---
## 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.

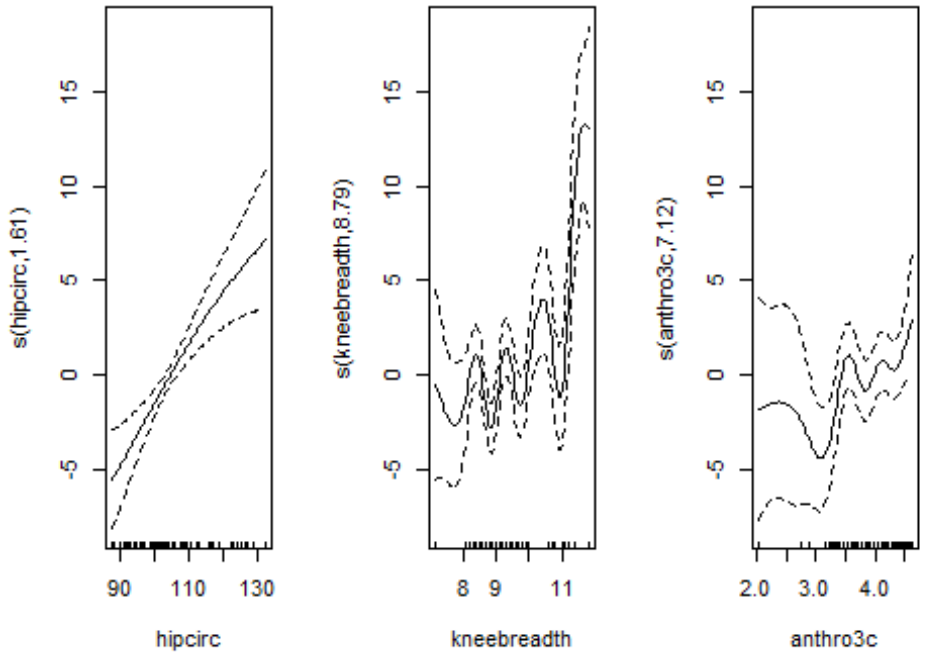
Answer:

The model indicates that all the predictor variables are significant at 0.5 significance level.

```
##
## 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
```

Answer:

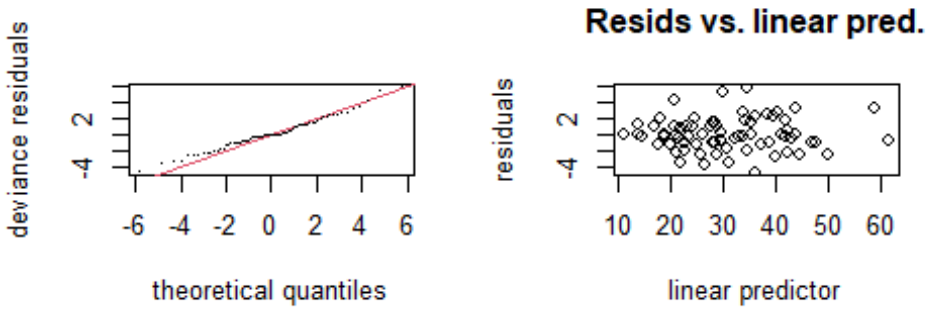
The plots shows the predictors that were smoothed for the model. The variables that were included in this model are the same variables that we concluded that they needed smoothing.

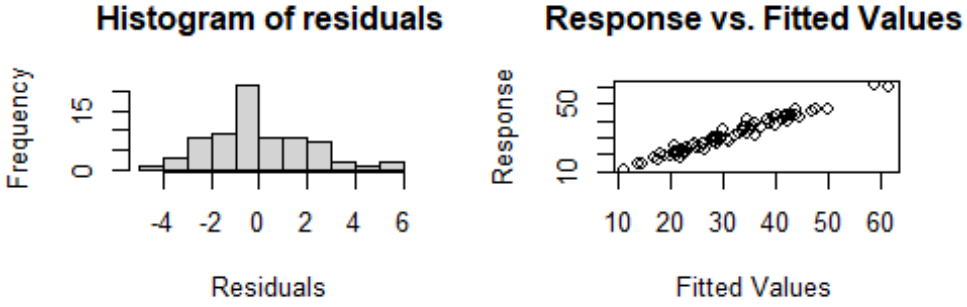



Answer:

The GGv and the AIC are a bit lower than the previous model and the R^2 is alomost the same as the last model indicating that this model performs slighlty better than the last model.

```
## GGV

## [1] 7.946447

## AIC

## [1] 343.2562

## Adjusted R^2

## [1] 0.9536683

## DF

## [1] 17.52001
```



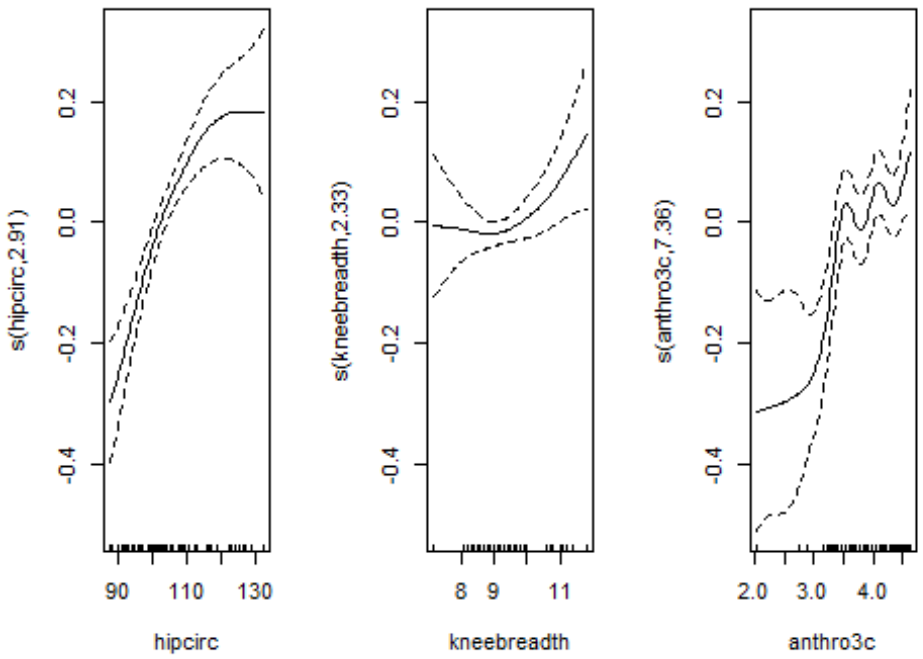




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 24 iterations.
## The RMS GCV score gradient at convergence was 0.0001386163 .
## The Hessian was positive definite.
## Model rank =  30 / 30
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## 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    0.44
## s(kneebreadth) 9.00 8.79    1.06    0.71
## s(anthro3c)    9.00 7.12    0.91    0.16
```

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).

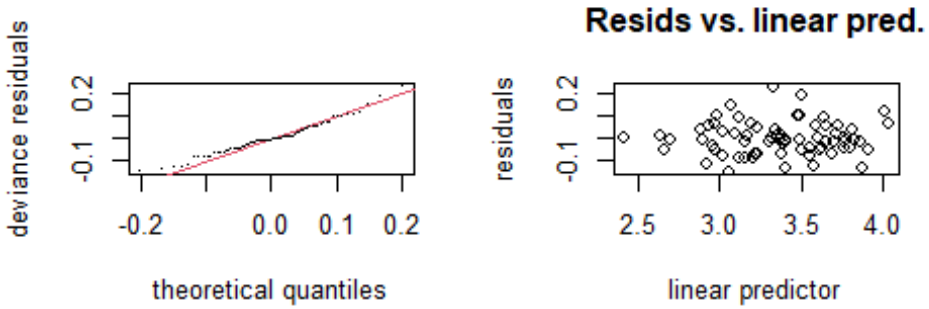
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(DEXfat) ~ waistcirc + s(hipcirc) + s(kneebreadth) + anthro3a +
##     s(anthro3c)
##
```

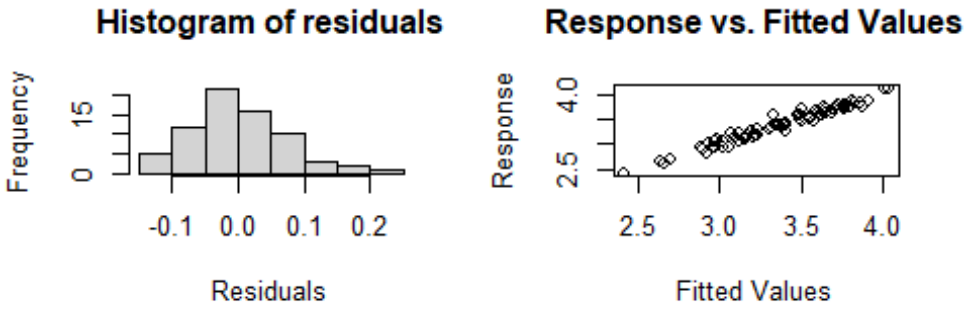
```
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.139779   0.237083   9.025  1.8e-12 ***
## waistcirc   0.004418   0.001806   2.447 0.017610 *
## anthro3a    0.215488   0.054600   3.947 0.000226 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 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
```




```
## GGV

## [1] 0.008813659

## AIC

## [1] -136.47

## Adjusted R^2
```

```
## [1] 0.9522733

## DF

## [1] 12.59274
```




```
##
## 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   0.085 .
## s(kneebreadth) 9.00 2.33    0.83   0.060 .
## s(anthro3c)    9.00 7.36    0.99   0.405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion:

The smoothed predictors are highly significant at 0.5 level except for *kneebreadth* which has p-value that is above 0.5. All the parametric coefficients are ignificant at 0.5 level as

well. By observing the smoothed terms they seem to be more smooth than terms in part c and b.

The GGv for this model is **GGV = 0.008813659** and the AIC is **AIC = -136.47** They are both tremendously lower than the GGV and AIC for models in part b and part c. The R^2 for this model **0.9522733** which is lower than the R^2 for the other models. This model explains does explain the variation in the response variable better than the other two models. The Df for this model is the lowest out of all with **DF= 12.59**. The diagnostics plots seem to be the similar for all models. This models residuals are distributed randomly around zero and both histogram and qq plots seem to have a normal distribution. I think the Part c model is the most appropriate model since it explains the highest percentage of variation in the response variable **0.9537**.

```
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?

##
##   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
```

Answer:

Variable age was the only variable removed by this selection method.

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.

#Overview:

The GlaucomaM data has 196 observations in two classes. 62 variables are derived from a confocal laser scanning image of the optic nerve head, describing its morphology. Observations are from normal and glaucomatous eyes, respectively.

#Data and Model:

All variables are derived from a laser scanning image of the eye background taken by the Heidelberg Retina Tomograph. Most of the variables describe either the area or volume in certain parts of the papilla and are measured in four sectors (temporal, superior, nasal and inferior) as well as for the whole papilla (global). The global measurement is, roughly, the sum of the measurements taken in the four sector.

We will fit an additive model to the data where class will be **Class** as response variable. We will use the **gamboost** to select the most influential variables since we a big number of covarites.

#Results

We fit a model and we selected the variables with most impact on the response variable. The data frame below lists the selected variables and their influence on the probability of suffering from glaucoma. The variable with the highest probability is variable **as** with probability of **0.17** and the variable with lowest probability is variable **mv** with probability of **0.01**

```
##      Probablities
## as           0.17
## abrs         0.11
## hic          0.11
## mhcg         0.10
## mhcn         0.08
## mhci         0.08
## phcg         0.07
## phcn         0.06
## phci         0.04
## hvc          0.03
## vass         0.03
## vars         0.03
## vari         0.03
## mdn          0.02
## mdi          0.01
## tms          0.01
```

```
## tmi          0.01
## mv           0.01
```

strangely, All the selected explanatory variables are insignificant at level of 0.5. All the values have estimated degrees of freedom of 1 except for **as,mhcg,mhci,phcg,hvc,vari,mdi,and tmi**. which have EDF above 1. The **Deviance explained = 100% and R-sq.(adj) = 1* might be a bit suspicious and indicate that the model might be be fitting the data too well.

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Class ~ s(as) + s(abrs) + s(hic) + s(mhcg) + s(mhcn) + s(mhci) +
##     s(phcg) + s(phcn) + phci + s(hvc) + s(vass) + s(vars) + s(vari) +
##     s(mdn) + s(mdi) + s(tms) + s(tmi) + s(mv)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -17.56    1612.30  -0.011    0.991
## phci         -265.43   15824.55  -0.017    0.987
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(as)   1.868  1.922  0.002   0.999
## s(abrs) 1.000  1.000  0.000   0.985
## s(hic)  1.000  1.000  0.001   0.977
## s(mhcg) 1.362  1.389  0.000   1.000
## s(mhcn) 1.000  1.000  0.000   0.991
## s(mhci) 2.156  2.206  0.001   1.000
## s(phcg) 4.067  4.131  0.001   1.000
## s(phcn) 1.000  1.000  0.000   0.997
## s(hvc)  4.490  4.566  0.002   1.000
## s(vass) 1.000  1.000  0.003   0.959
## s(vars) 1.000  1.000  0.000   0.996
## s(vari) 2.169  2.230  0.001   1.000
## s(mdn)  1.000  1.000  0.001   0.971
## s(mdi)  1.339  1.360  0.001   1.000
## s(tms)  1.000  1.000  0.000   0.994
## s(tmi)  5.141  5.231  0.004   1.000
## s(mv)   1.000  1.000  0.000   0.985
##
## R-sq.(adj) =      1   Deviance explained =  100%
## UBRE = -0.65723  Scale est. = 1         n = 196
```

## Citation

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.

(n.d.). Retrieved November 06, 2020, from https://rpubs.com/kkuipers/529708