Homework 6

Snigdha Peddi

References

(PLEASE CLICK ON THE REFERENCE TO BE DIRECTED TO THE WEBPAGE)

Reference 1:Lecture Code(Chapter10Rcode.R), Reference 2, Reference 3, Reference 4, Reference 5, Reference 6,Reference 7

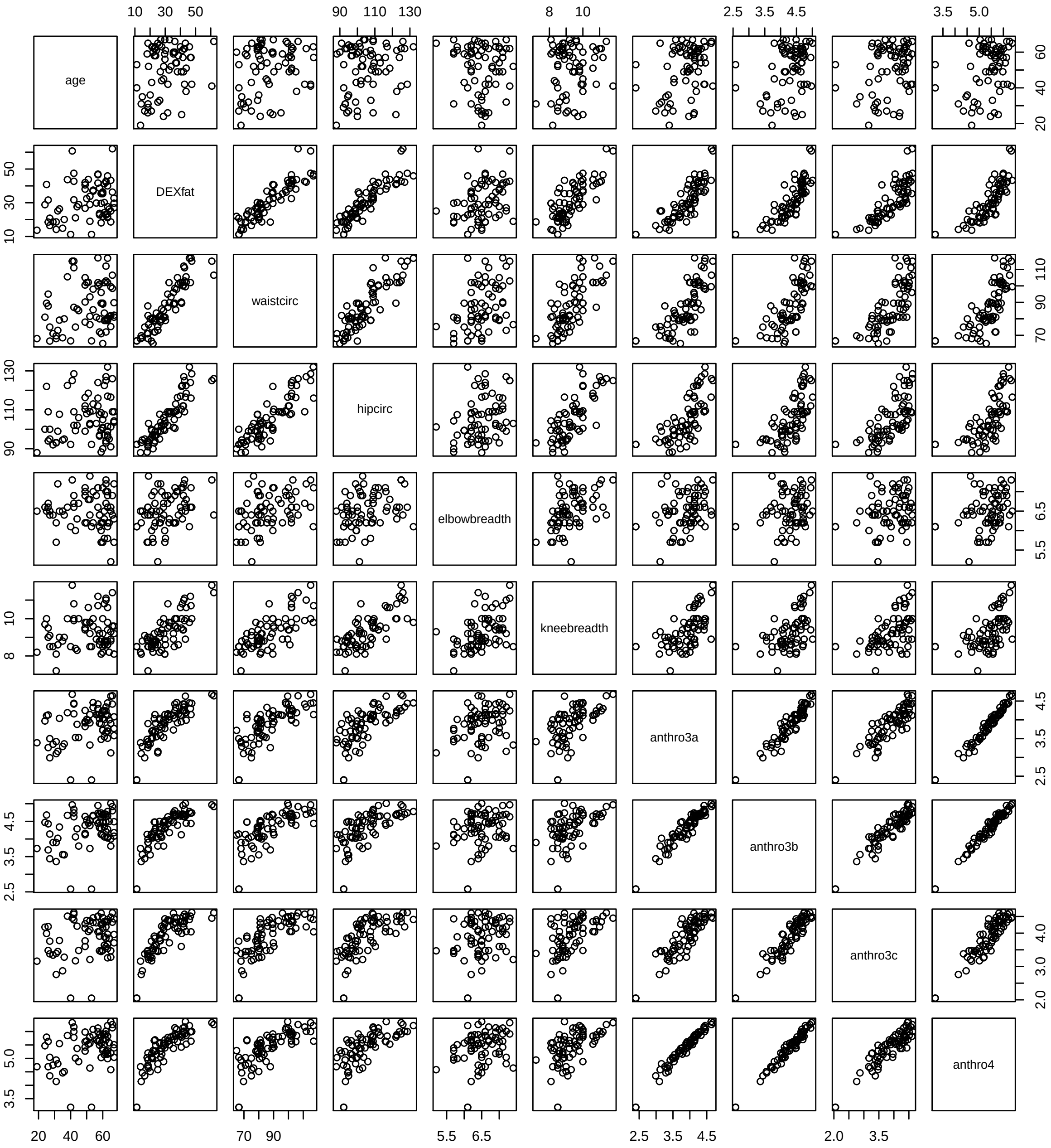
Exercises

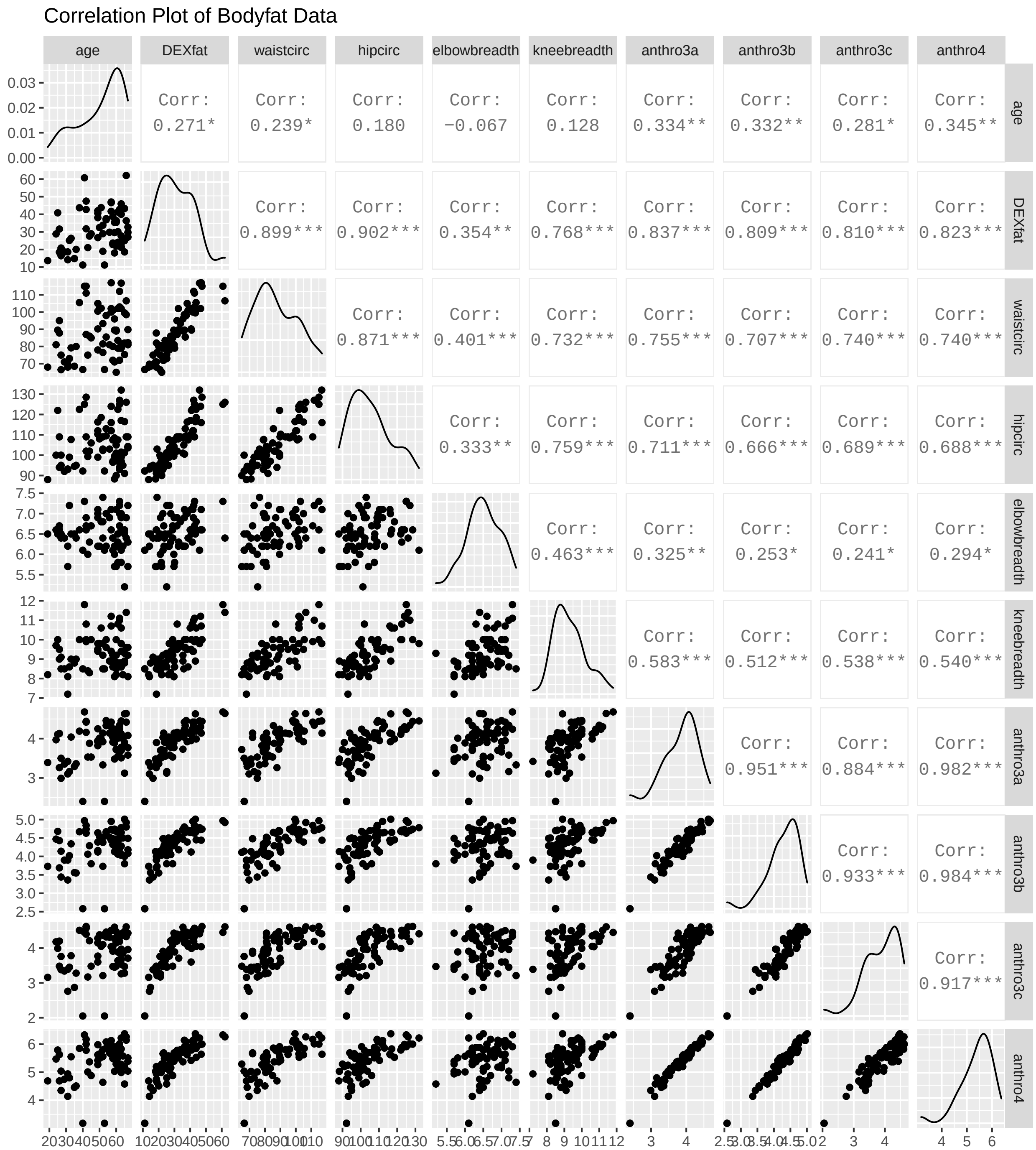
Question 1. (Ex. 10.1 pg 207 in HSAUR, modified for clarity) Consider the **bodyfat** data from the **TH.data** package introduced in Chapter 9.

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 1.a: Both the plots clearly show the relationship between DEXfat and other variables. A high correlation of greater than 0.8 between DEXfat and variables waistcirc,hipcirc,knessbredth,sum of three anthropometric measurements indicates that these variables are highly correlated and should be used in the model to predict the body fat.

Pairs Plot for Bodyfat Data





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

```
bodyfat_gam <- gam(DEXfat ~ s(age) + s(waistcirc) + s(hipcirc) +
s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) +
s(anthro3c), data = bodyfat)
```

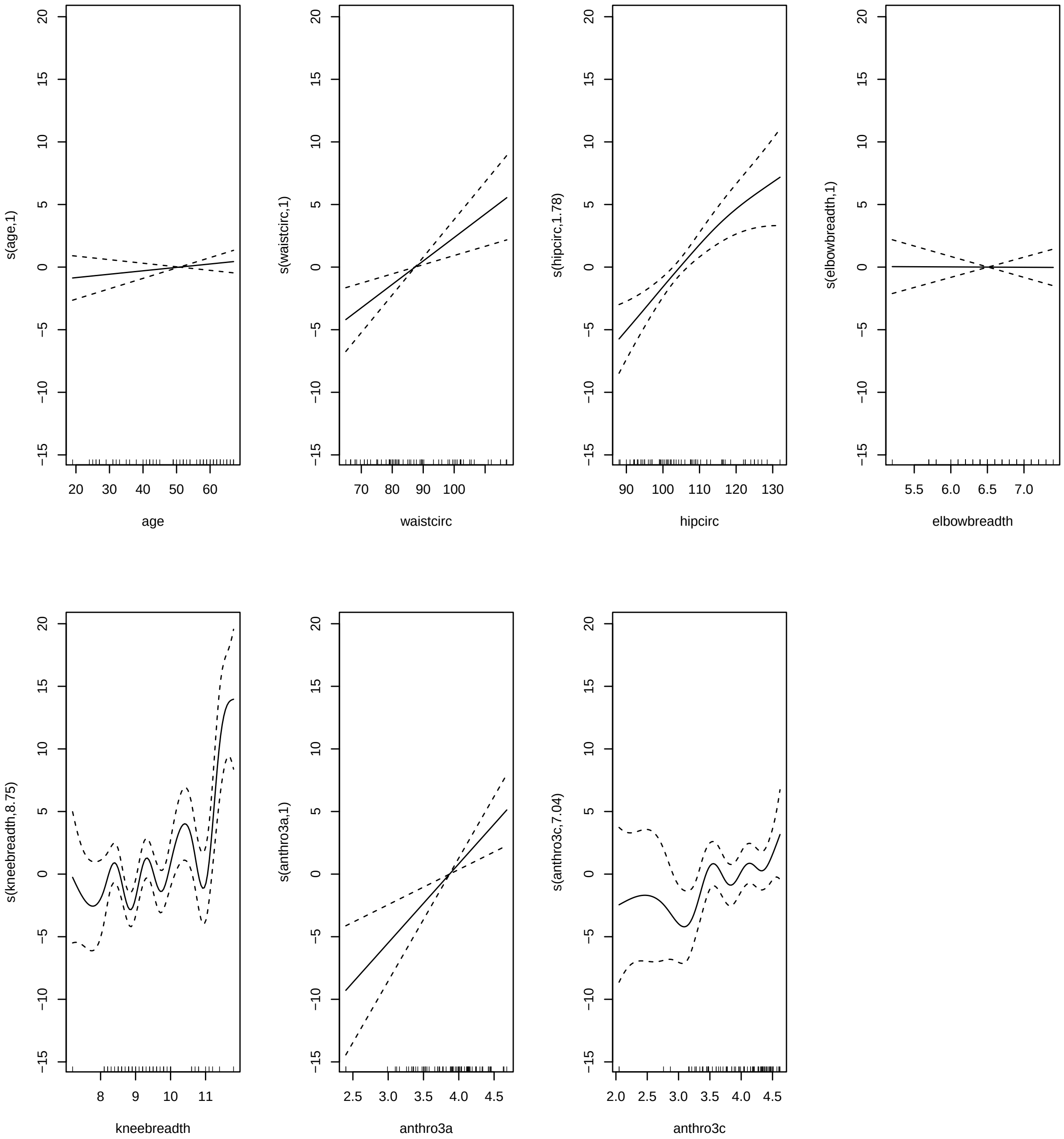
- Assess the **summary()** and **plot()** of the model (don't need GGPLOT for a plot of the model). Are all covariates informative? Should all covariates be smoothed or should some be included as a linear effect?

Answer 1.b: A Generalized Linear regression model is fit.

The Summary of the model shows that the variables age,elbowbreadth and anthro3c are not significant at 95% confidence interval.The plots show that age, waistcirc, elbowbreadth and anthro3a are linearly related and also is clear from the summary that these variables have a degree of freedom 1. The model with a linear fit of waistcirc shows that the variable is still significant at 95% confidence interval. Another model built with linear fit of variables age,elbowbreadth and anthro3c shows that anthro3c is significant at 95% confidence interval. The final model with variables waistcirc and anthro3c fit linearly shows that these variables are significant.

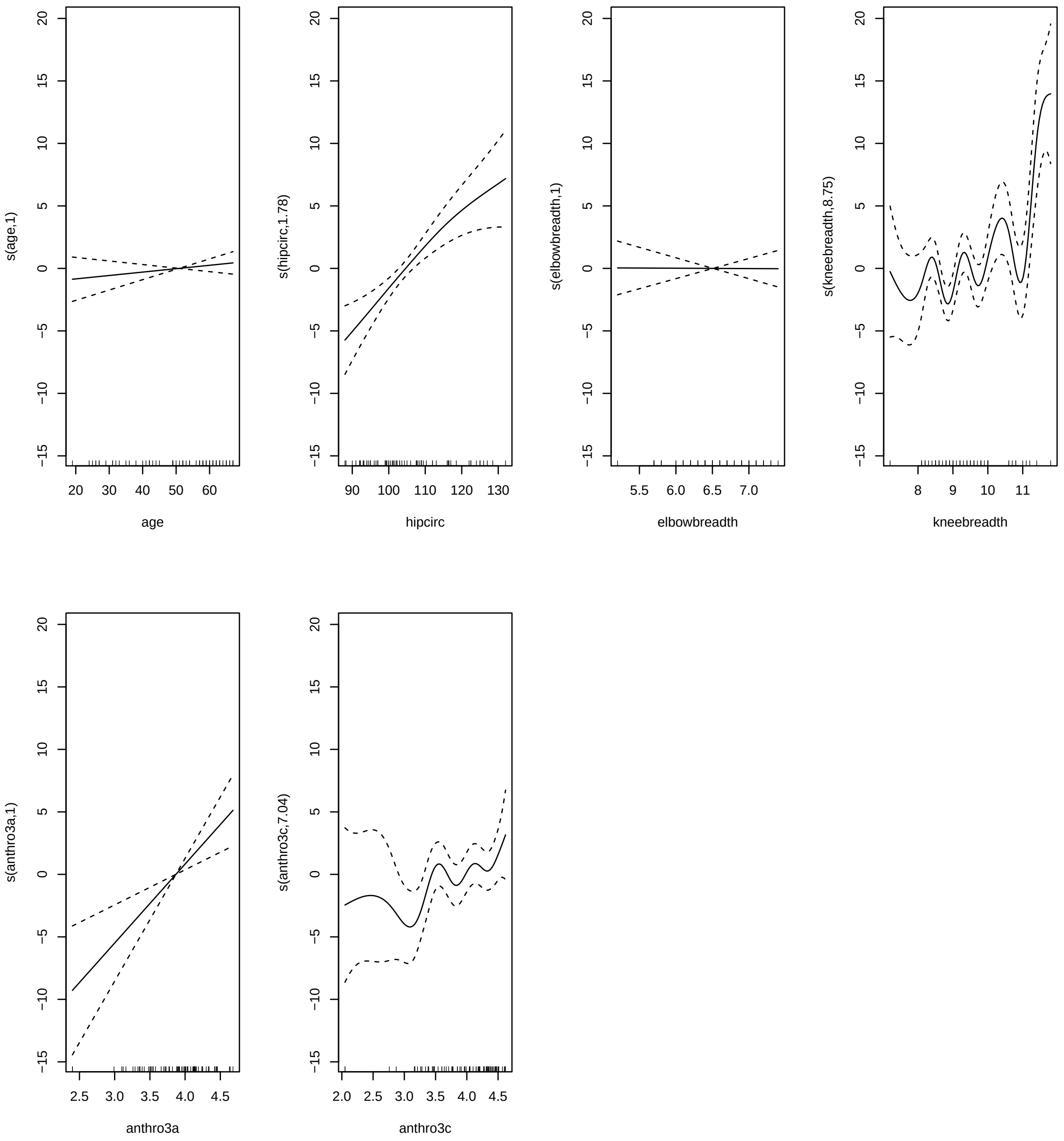
Comparing all the models variable waistcirc can be included as a linear effect. This can be confirmed by comparing the AIC of the 4 models.The initial model and model with linearly fit waistcirc has similar and lower AIC and are the best models among the 4 models fit.

```
##
## 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 0.332964
## s(waistcirc)    1.000  1.000 10.821 0.001844 **
## s(hipcirc)      1.775  2.235  9.917 0.000152 ***
## s(elbowbreadth) 1.000  1.000  0.001 0.972242
## s(kneebreadth)  8.754  8.960  6.180 3.59e-06 ***
## s(anthro3a)     1.000  1.000 12.966 0.000725 ***
## s(anthro3c)     7.042  8.041  1.798 0.100242
## ---
## 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
```



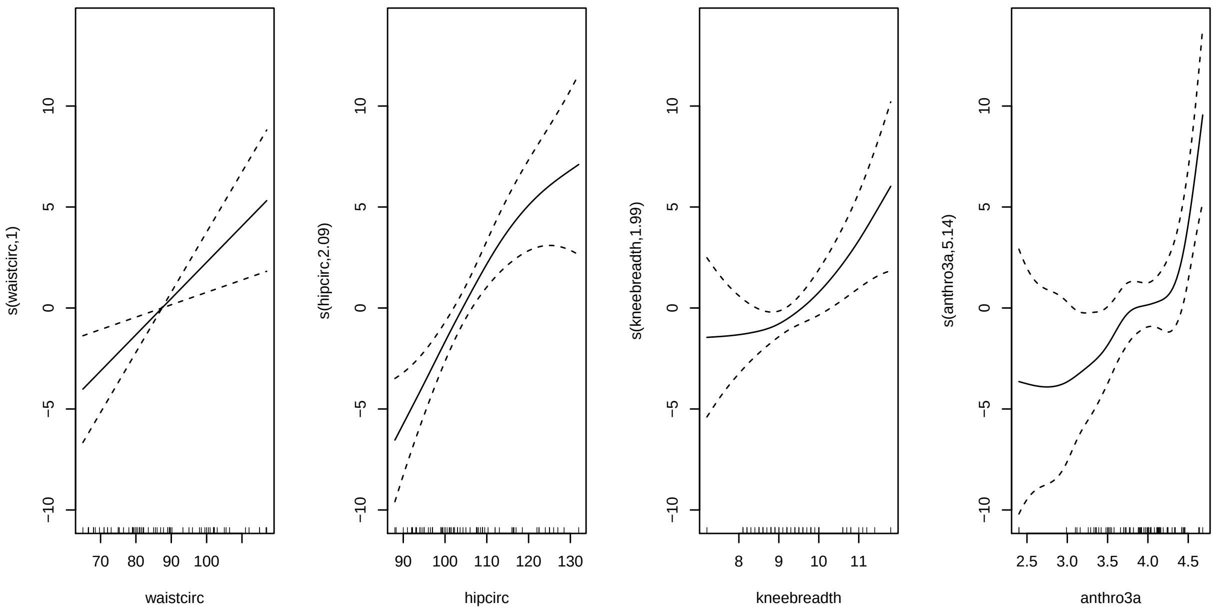
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## DEXfat ~ s(age) + waistcirc + s(hipcirc) + s(elbowbreadth) +
##     s(kneebreadth) + s(anthro3a) + s(anthro3c)
##
## Parametric coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.42025    4.98232   2.894  0.00568 **
## waistcirc    0.18725    0.05692   3.289  0.00188 **
## ---
## 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 0.332964
## s(hipcirc)      1.775  2.235  9.917 0.000152 ***
## s(elbowbreadth) 1.000  1.000  0.001 0.972242
## s(kneebreadth)  8.754  8.960  6.180 3.59e-06 ***
## s(anthro3a)     1.000  1.000 12.966 0.000725 ***
## s(anthro3c)     7.042  8.041  1.798 0.100242
## ---
## 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
```



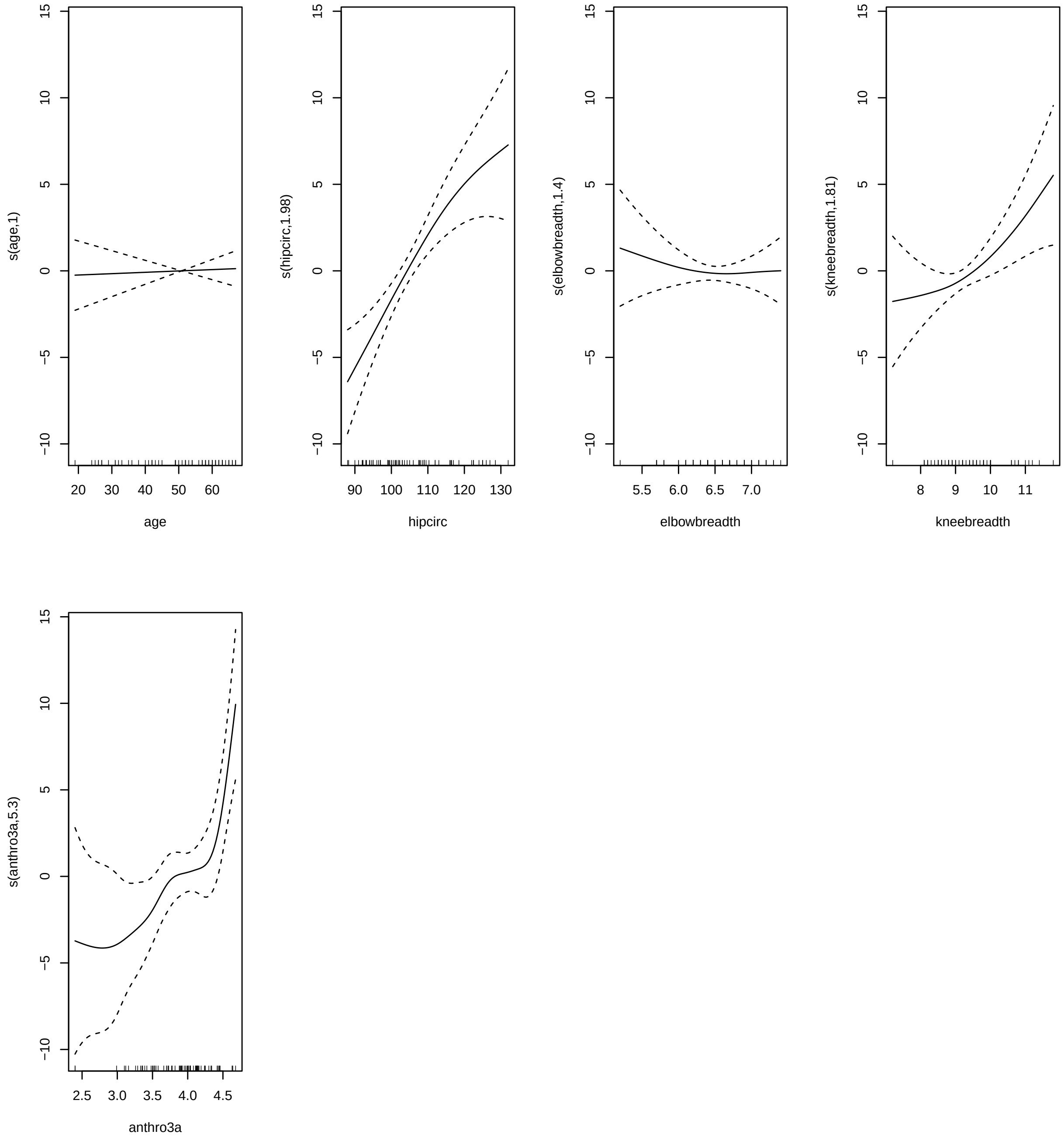
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## DEXfat ~ age + s(waistcirc) + s(hipcirc) + elbowbreadth + s(kneebreadth) +
##     s(anthro3a) + anthro3c
##
## Parametric coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.28433    9.99352   1.930   0.0586 .
## age           0.01489    0.03076   0.484   0.6301
## elbowbreadth -0.34717    0.93394  -0.372   0.7115
## anthro3c      3.34582    1.55970   2.145   0.0362 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df     F  p-value
## s(waistcirc)   1.000  1.000 9.237 0.003547 **
## s(hipcirc)     2.095  2.635 8.785 0.000145 ***
## s(kneebreadth) 1.985  2.497 3.572 0.026204 *
## s(anthro3a)    5.137  6.104 3.440 0.005239 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.933   Deviance explained = 94.5%
## GCV =  10.27  Scale est. = 8.2138    n = 71
```



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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.91758    6.87778   0.279  0.78141
## waistcirc    0.18244    0.05889   3.098  0.00304 **
## anthro3c     3.32587    1.54859   2.148  0.03604 *
## ---
## 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 0.061 0.806559
## s(hipcirc)      1.982  2.488 9.011 0.000162 ***
## s(elbowbreadth) 1.399  1.693 0.465 0.683226
## s(kneebreadth)  1.813  2.263 3.530 0.030944 *
## s(anthro3a)     5.304  6.290 3.630 0.003353 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## R-sq.(adj) =  0.933   Deviance explained = 94.6%
## GCV = 10.233  Scale est. = 8.1434    n = 71
```



```
##   bodyfat_gam.aic bodyfat_gam2.aic bodyfat_gam3.aic bodyfat_gam4.aic
## 1         345.708          345.708         365.5719         365.1702
```

- 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.

The total degrees of freedom is equal to number of terms/variables in the model.However, flexible regression models have nonlinear parameters that require more than one df. For this model the edf(effective degrees of freedom) is 21 indicating that few terms are smoothed ,increasing the number of parameters.The GCV score or the generalized cross validation score is an estimate of mean square error of the LOOCV process. A moderate GCV score along with high AIC score indicates that the model performance can be improved by variable selection.

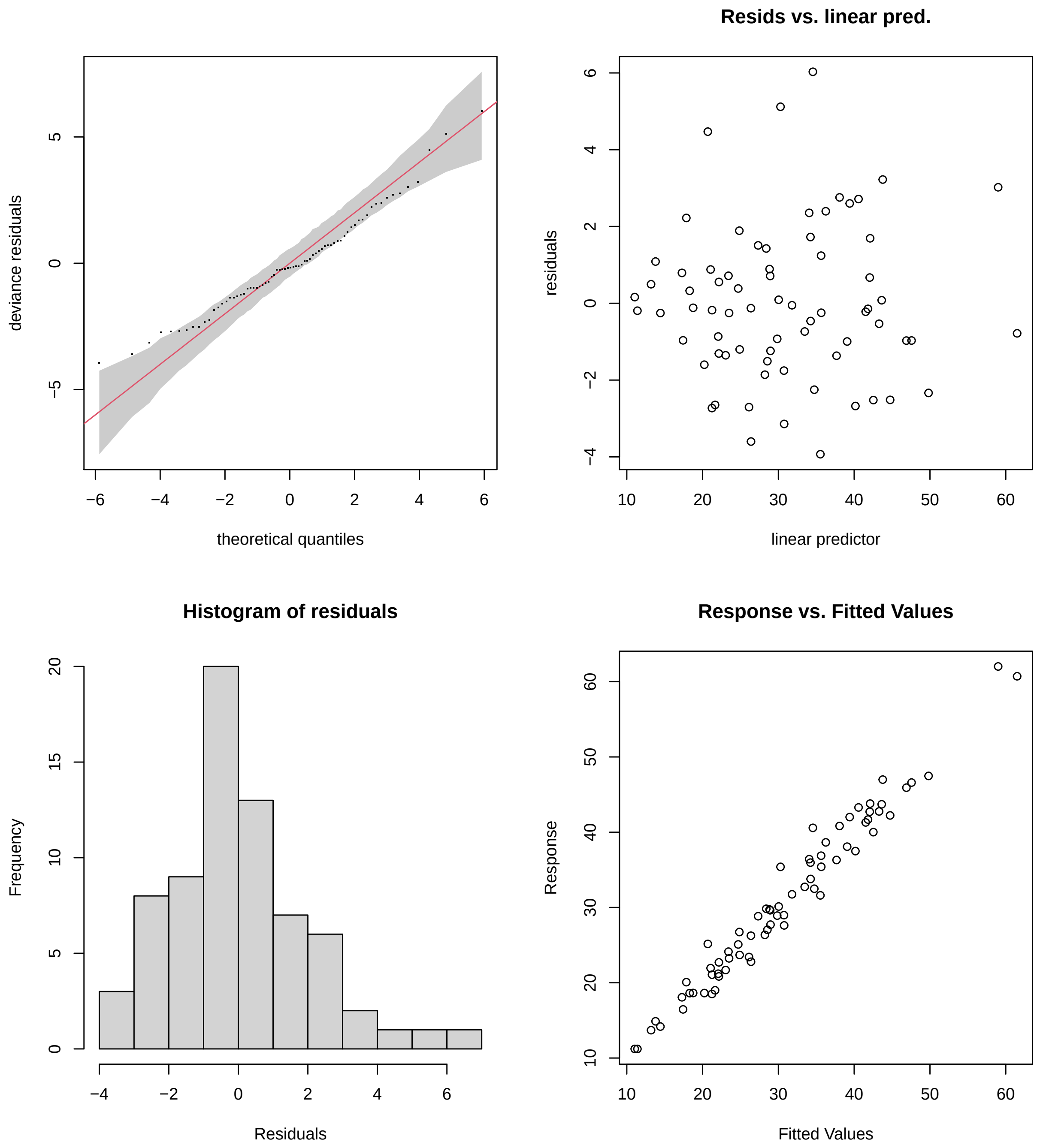
```
## AIC of model 1: 345.708
```

```
## GCV of model 1: 8.435412
```

```
## Degrees of Freedom of model 1: 21.57091
```

- Produce a diagnostic plot using **gam.check()** function. Are any concerns raised by the diagnostic plot?

The Quantile plot show that few points do not fall within the confidence interval.The residual vs.linear predictor plot shows that the spread widening out which is of concern.The histogram looks guassian and right skewed.The Response vs.Fitted values plot have linear fit indicating the model is performing good.The model can be improved by variable selection or using high order covariates.



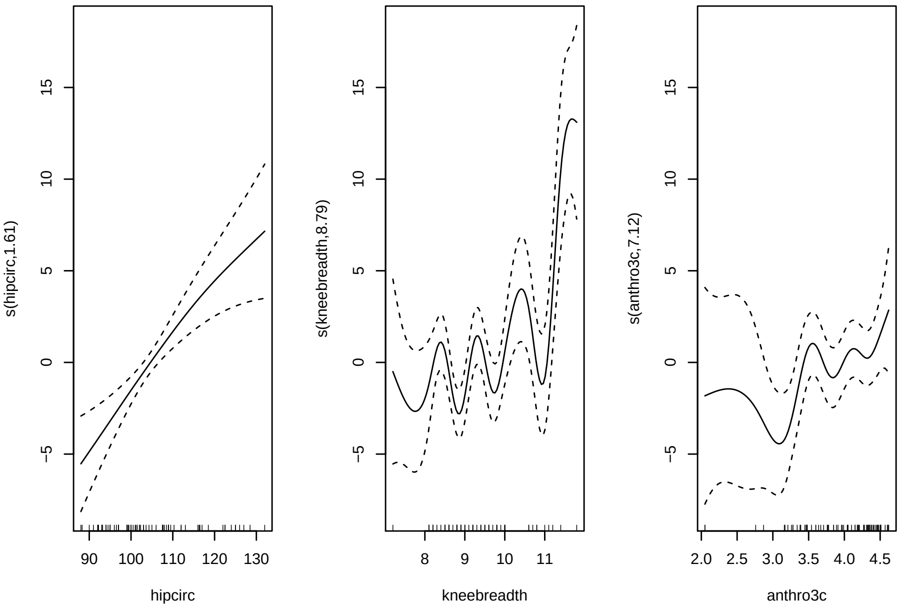
```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 41 iterations.
## The RMS GCV score gradient at convergence was 2.767255e-07 .
## The Hessian was positive definite.
## Model rank =  64 / 64
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## 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    0.05 *
## s(waistcirc)    9.00 1.00    0.94    0.34
## s(hipcirc)      9.00 1.78    1.02    0.52
## s(elbowbreadth) 9.00 1.00    0.81    0.05 *
## s(kneebreadth)  9.00 8.75    1.08    0.76
## s(anthro3a)     9.00 1.00    1.09    0.78
## s(anthro3c)     9.00 7.04    0.89    0.13
## ---
## 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.

```
  bodyfat_gam2 <- gam(DEXfat~ waistcirc + s(hipcirc) +
  s(kneebreadth)+ anthro3a +
  s(anthro3c), data = bodyfat)
```

```
##
## 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 0.000103 ***
## s(kneebreadth) 8.793  8.970  6.780 2.48e-06 ***
## s(anthro3c)    7.117  8.103  2.126 0.048737 *
## ---
## 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
```



```
## AIC of model 5: 343.2562
```

```
## GCV of model 5: 7.946447
```

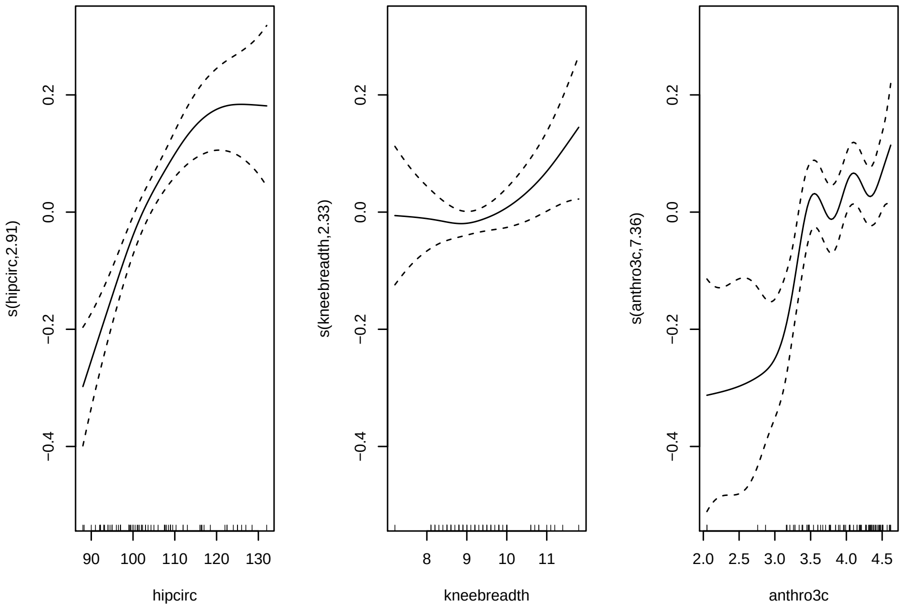
d) Again fit an additive model to the body fat data, but this time for a log-transformed response. Compare the three models, which one is more appropriate? (Hint: use AIC, GCV, residual plots, etc. to compare models).

Answer 1.d: An additive model is fit with the log-transformed DEXfat variable.By comparing the 3 models, the lower AIC and GCV scores clearly indicate that model with log transformed response variable and linearly fit waistecirc and anthro3a variables(Model 3) has better performance.

```
##
## 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  8.8e-07 ***
## s(kneebreadth) 2.325  2.962  2.027 0.128320
## s(anthro3c)    7.358  8.263  4.678 0.000144 ***
## ---
## 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
```

```
##
##
##  AIC of model 6: -136.47
```

```
## GCV of model 6: 0.008813659
```



```
##   AIC.mod6 AIC.mod5 AIC.mod
## 1  -136.47 343.2562 345.708
```

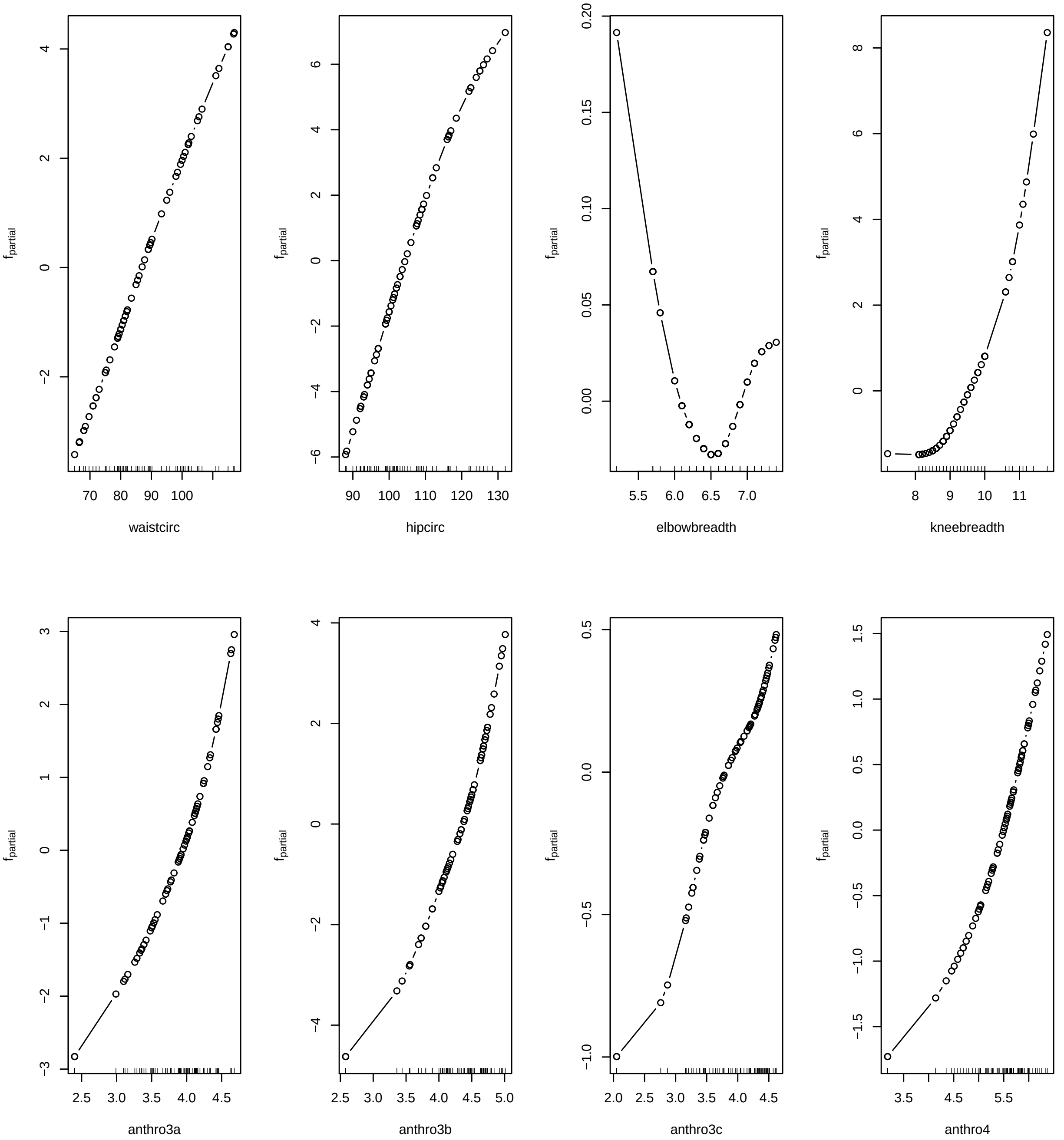
```
##           GCV.mod6 GCV.mod5  GCV.mod
## GCV.Cp 0.008813659 7.946447 8.435412
```

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?

```
bodyfat_boost <- gamboost(DEXfat~., data = bodyfat)
bodyfat_aic <- AIC(bodyfat_boost)
bf_gam <- bodyfat_boost[mstop(bodyfat_aic)]
```

Answer 1.e: The AIC suggests that the bossting algorithm should be stopped after 51 iterations.The Age variable is removed by the AIC method and the resulting AIC is very low indicating this model is performing better than the previous models.The plots of the models shows that most of the variables are smoothed enough to form have a linear relationship with the DEXfat.Kneebresdth slowly increased and shows a linear relation. Only the elbowbreadth has non linear relation decreasing to the most extent and slowly increasing thereafter.

```
##
##   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
```

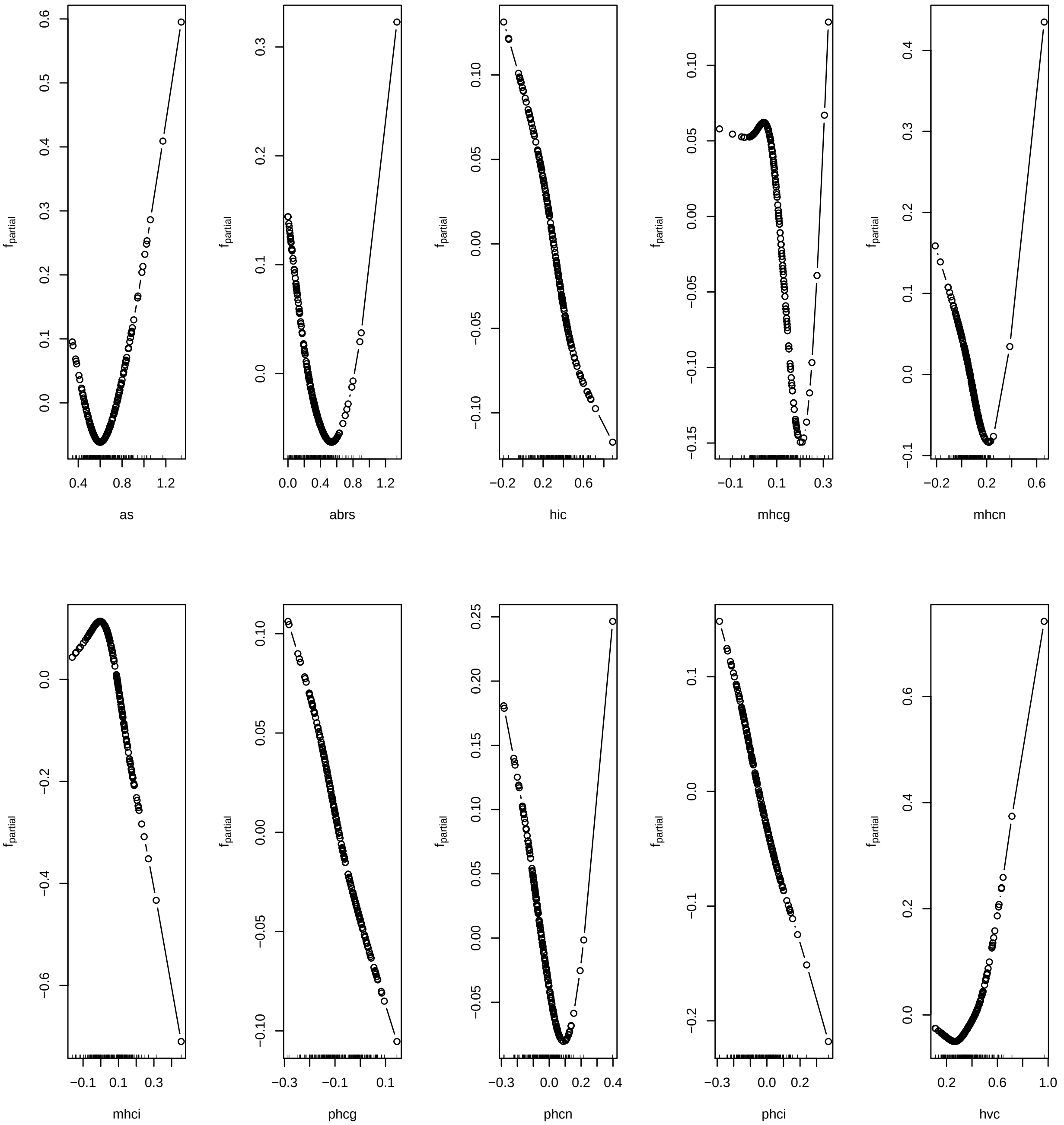


```
##
##
##  AIC of model: 3.268173
```

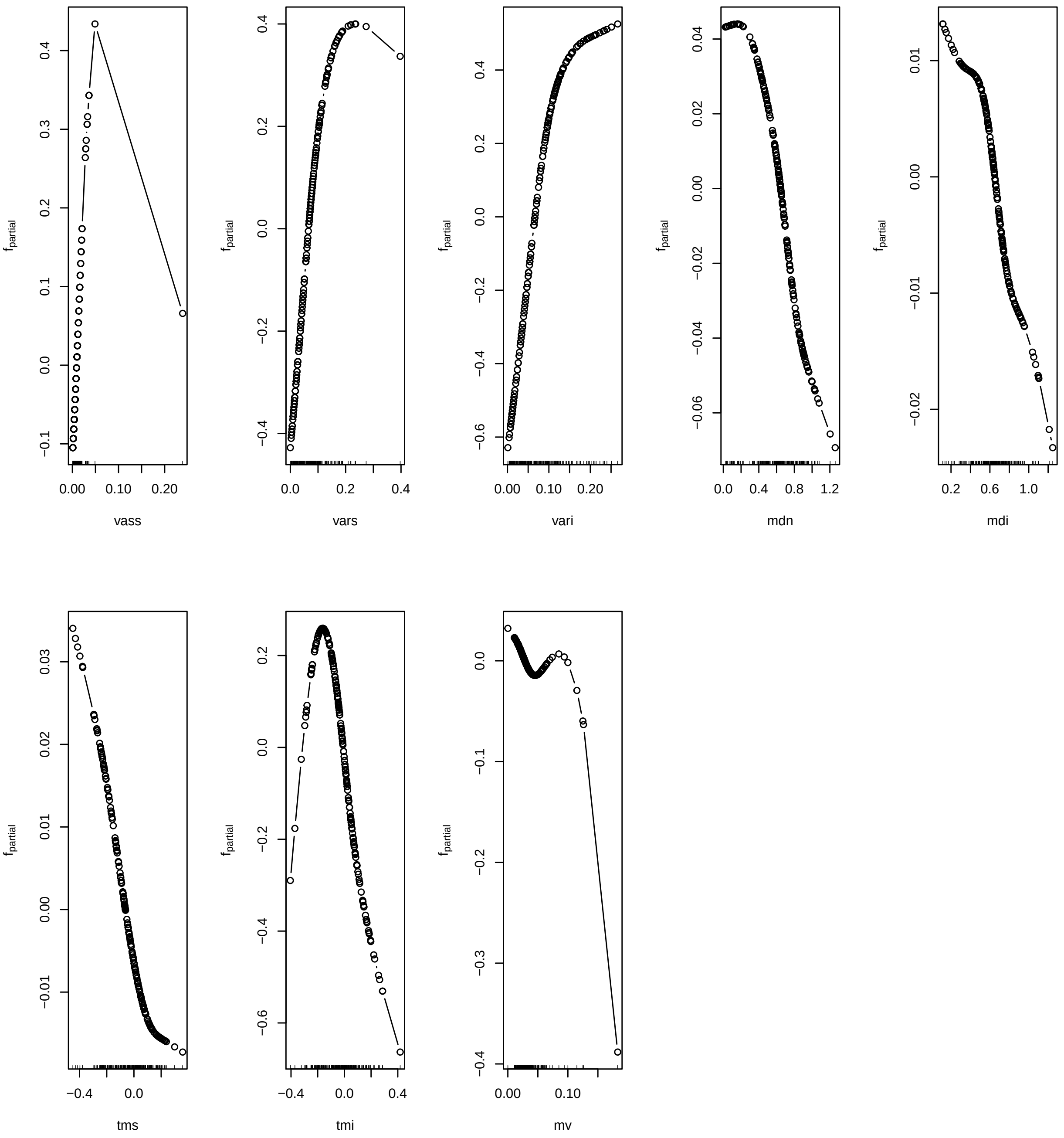
Question 2. (Ex. 10.3 pg 208 in HSAUR, modified for clarity) Fit an additive model to the **glaucomaM** data from the **TH.data** library with *Class* as the response variable. Read the description of the dataset and the goals of the experiment. Which covariates should be in the model and what is their 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.

Answer 2: The goal of the experiment to find a way to decide whether an eye is affected by glaucoma or not. As suggested, **gamboost()** is used to fit the Generalized Additive model as there are 63 covariates.AIC of the model suggests that the boosting algorithm is to be stopped at 100 iterations and has 62 base learners.The partial contribution of 18 important covariates out of 63 variables are shown in the plots.The plots shows that out of 18 covariates that are not linearly related, lower values of “volume above reference superior-vars” and “volume above reference inferior-vari” and higher values of other variables indicate presence of Glaucoma.

```
##
##   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
```



f_partial
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.4
0.8
1.2
as
f_partial
0.0
0.1
0.2
0.3
0.0
0.4
0.8
1.2
abrs
f_partial
−0.10
−0.05
0.00
0.05
0.10
−0.2
0.2
0.6
hic
f_partial
−0.15
−0.10
−0.05
0.00
0.05
0.10
−0.1
0.1
0.3
mhcg
f_partial
−0.1
0.0
0.1
0.2
0.3
0.4
−0.2
0.2
0.6
mhcn
f_partial
−0.6
−0.4
−0.2
0.0
−0.1
0.1
0.3
mhci
f_partial
−0.10
−0.05
0.00
0.05
0.10
−0.3
−0.1
0.1
phcg
f_partial
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
−0.3
0.0
0.2
0.4
phcn
f_partial
−0.2
−0.1
0.0
0.1
−0.3
0.0
0.2
phci
f_partial
0.0
0.2
0.4
0.6
0.2
0.6
1.0
hvc



AIC of the model:

```
## [1] 139.6983
## Optimal number of boosting iterations: 100
## Degrees of freedom (for mstop = 100): 10.10971
```