Analysing Longitudinal Data from HSAUR

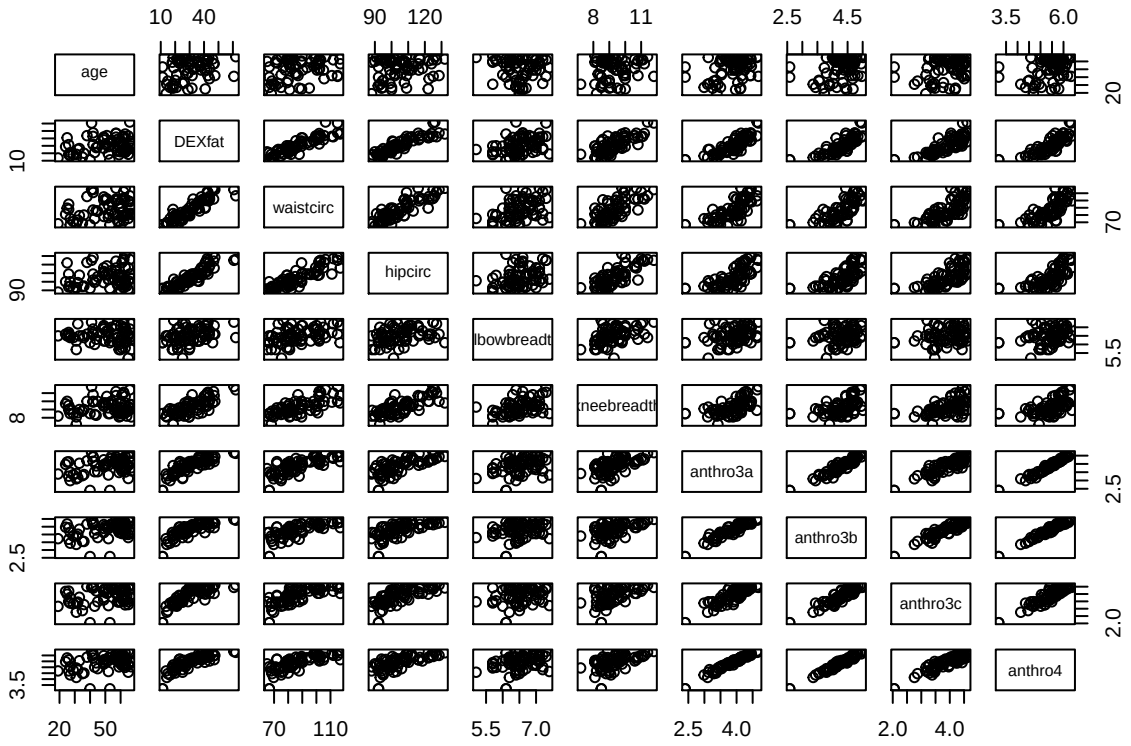
Yamuna Dhungana

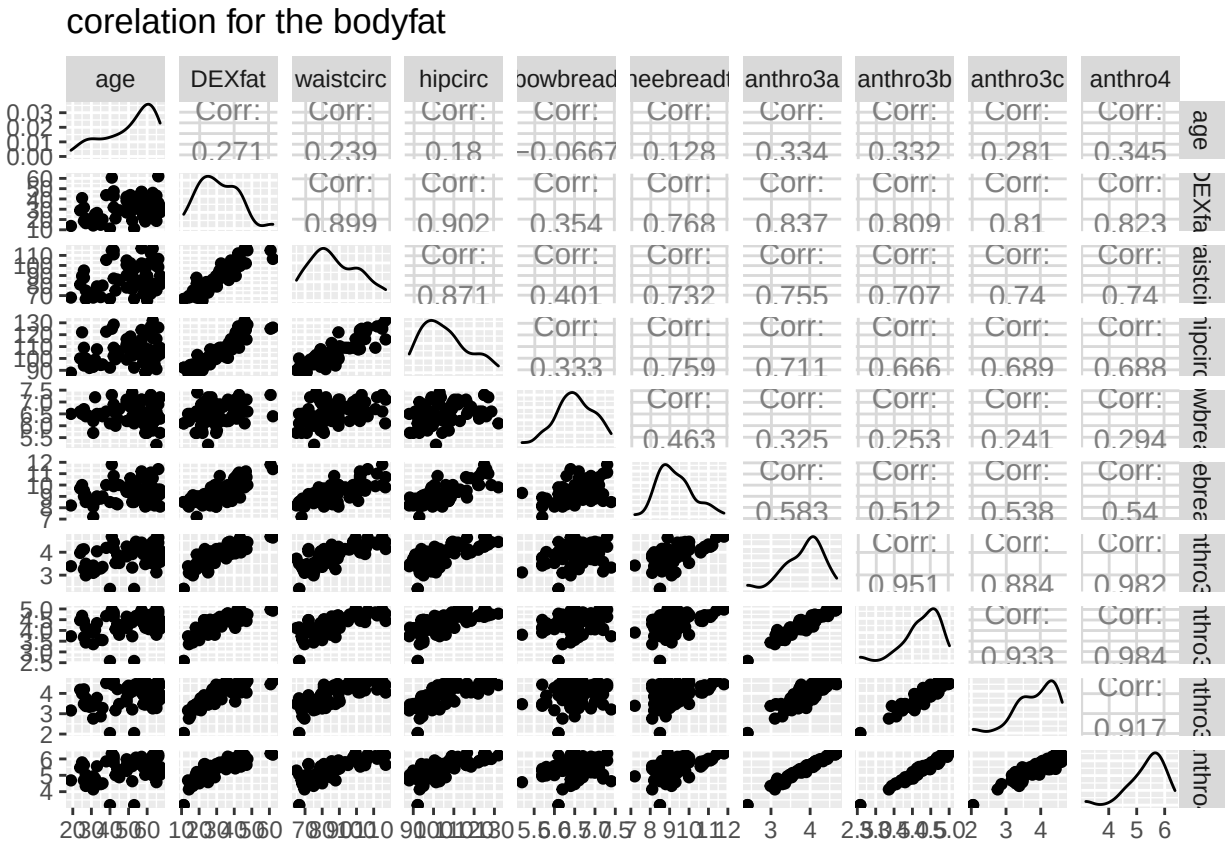
Exercises

Installing the required libraries.

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.





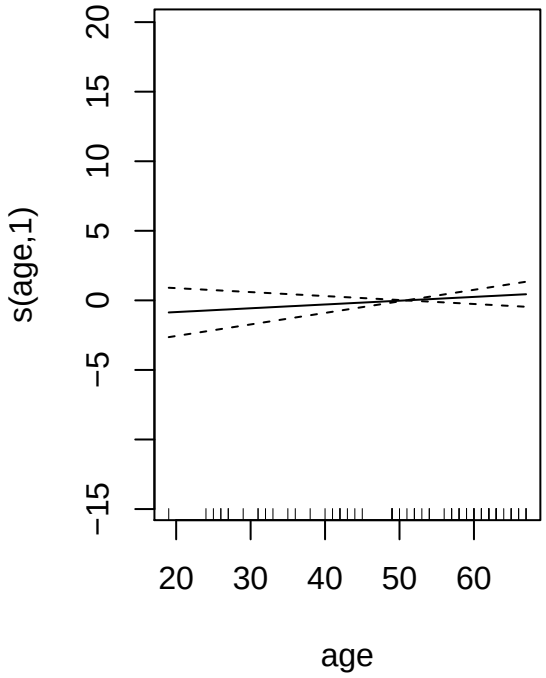
As we know, when the value of correlation is equal to 1, variables are highly correlated. The relation becomes weak when the value of correlation begins to decrease. From the normal plot function, we can say that there is some relationship between the variables, besides age and the elbow width.When using the gaally package, we can view the relationship between the variables with its correlation values. The correlation value shows that age and the elbowbreadth are not correlated with any of the other variables because the values of correlation are comparatively less. The graph shows some relation between them as well. The DEXfat correlation values are nearly equal to 1, and these values say that the variable is highly correlated. Hence, we can use it as a variable to predict the model. Besides DEXfat, hipcirc and waistcirc also can be used as the alternative predictor as well because of their higher correlation values.

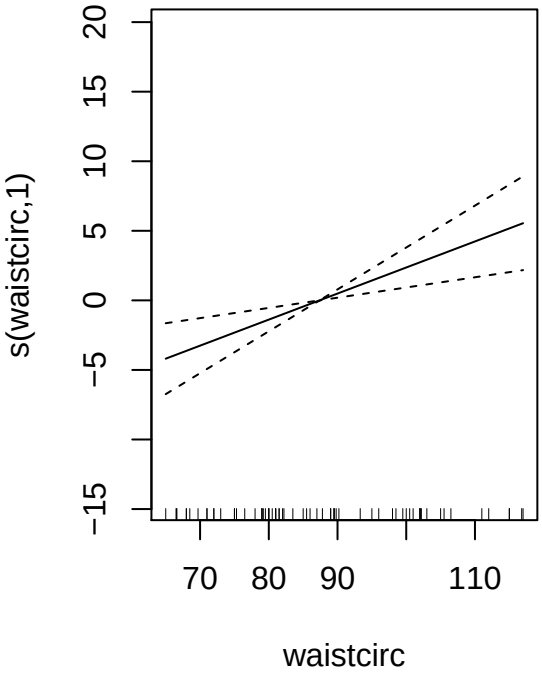
b) For feasability of the class, fit a generalised additive model assuming normal errors using the following code.

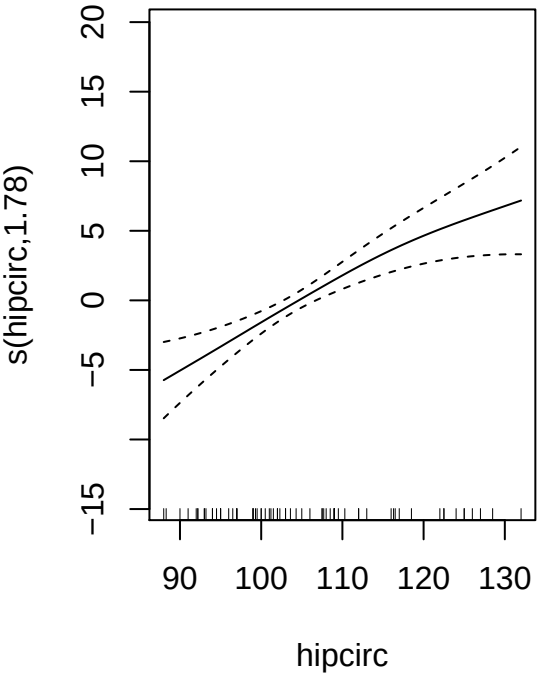
```
bodyfat_gam <- gam(DEXfat ~ s(age) + s(waistcirc) + s(hipcirc) +
        s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) +
        s(anthro3c), data = bodyfat)
```

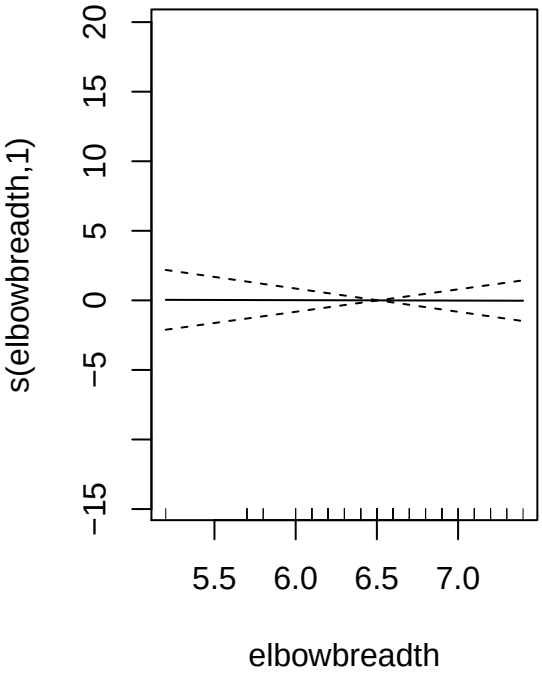
- Assess the **summary()** and **plot()** of the model (don't need GGPLOT for a plot of the model). Are all covariates informative? Should all covariates be smoothed or should some be included as a linear effect?
- Report GCV, AIC, and total model degrees of freedom. Discuss how certain you are that 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.

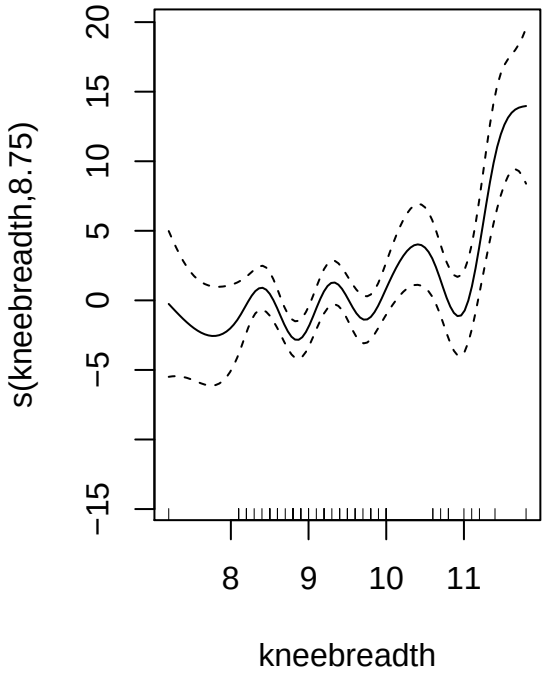
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## DEXfat ~ s(age) + s(waistcirc) + s(hipcirc) + s(elbowbreadth) +
##     s(kneebreadth) + s(anthro3a) + s(anthro3c)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.7828     0.2847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

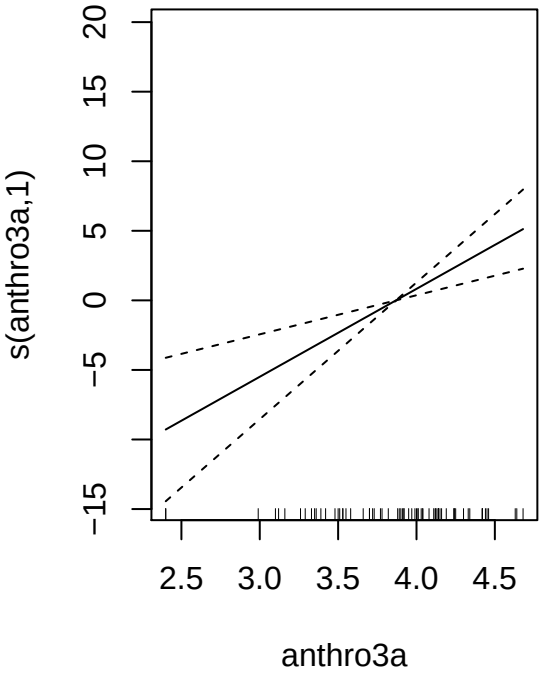

s(age,1)
20
15
10
5
0
−5
−15
20
30
40
50
60
age


s(waistcirc,1)
20
15
10
5
0
−5
−15
70
80
90
110
waistcirc

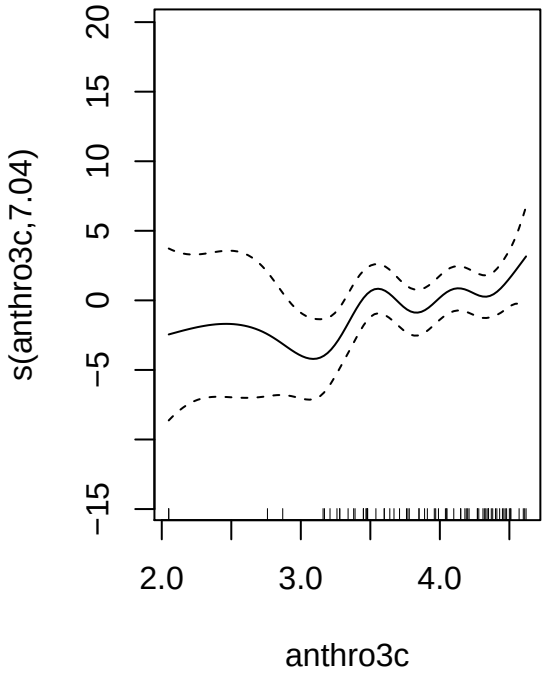

s(hipcirc,1.78)
20
15
10
5
0
−5
−15
90
100
110
120
130
hipcirc

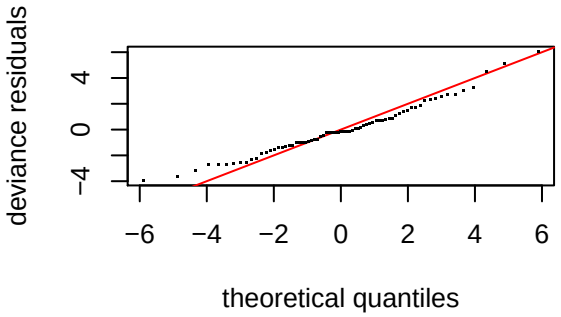

s(elbowbreadth,1)
20
15
10
5
0
−5
−15
5.5
6.0
6.5
7.0
elbowbreadth

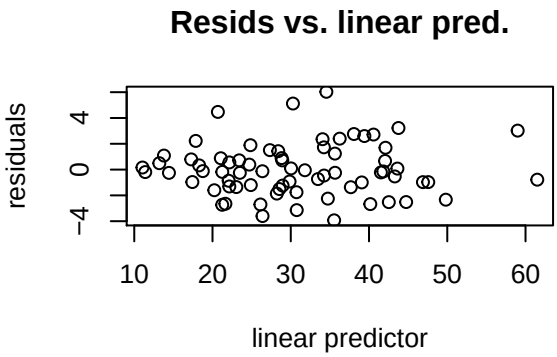


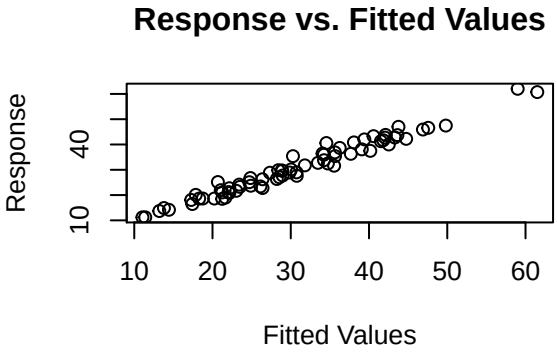



```
## Diagnostic plots
```


s(anthro3c,7.04)
20
15
10
5
0
−5
−15
2.0
3.0
4.0
anthro3c








```
##
## 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   0.045 *
## s(waistcirc)    9.00 1.00    0.94   0.300
## s(hipcirc)      9.00 1.78    1.02   0.570
## s(elbowbreadth) 9.00 1.00    0.81   0.030 *
## s(kneebreadth)  9.00 8.75    1.08   0.680
## s(anthro3a)     9.00 1.00    1.09   0.705
## s(anthro3c)     9.00 7.04    0.89   0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## GCA for the bodyfat data: 8.435412

## AIC for the bodyfat data:  345.708

## Degree of freedom for the bodyfat data:

##       (Intercept)          s(age).1          s(age).2          s(age).3
```

```
##      1.000000e+00      1.192818e-08     -9.080425e-09      2.149113e-09
##          s(age).4          s(age).5          s(age).6          s(age).7
##     -1.208855e-08      1.641067e-09     -9.583988e-09      7.304401e-10
##          s(age).8          s(age).9    s(waistcirc).1    s(waistcirc).2
##      7.508976e-08      1.000000e+00      1.026407e-08     -7.826376e-09
##    s(waistcirc).3    s(waistcirc).4    s(waistcirc).5    s(waistcirc).6
##      1.956036e-10     -1.098228e-08     -2.003051e-09     -1.213622e-08
##    s(waistcirc).7    s(waistcirc).8    s(waistcirc).9      s(hipcirc).1
##     -1.728571e-11      8.599913e-08      1.000000e+00      1.586383e-01
##      s(hipcirc).2      s(hipcirc).3      s(hipcirc).4      s(hipcirc).5
##     -4.823661e-02      1.188232e-02     -1.085696e-01      6.467778e-03
##      s(hipcirc).6      s(hipcirc).7      s(hipcirc).8      s(hipcirc).9
##     -1.779897e-01      2.196233e-03      9.308495e-01      1.000000e+00
## s(elbowbreadth).1 s(elbowbreadth).2 s(elbowbreadth).3 s(elbowbreadth).4
##      6.678706e-08     -1.947304e-08      1.326436e-08     -6.380128e-08
## s(elbowbreadth).5 s(elbowbreadth).6 s(elbowbreadth).7 s(elbowbreadth).8
##     -2.251820e-09     -7.581599e-08      4.338440e-10      5.079844e-07
## s(elbowbreadth).9  s(kneebreadth).1  s(kneebreadth).2  s(kneebreadth).3
##      1.000000e+00      9.991441e-01      1.016523e+00      9.948250e-01
##  s(kneebreadth).4  s(kneebreadth).5  s(kneebreadth).6  s(kneebreadth).7
##      9.693455e-01      9.646995e-01      8.956023e-01      7.945859e-01
##  s(kneebreadth).8  s(kneebreadth).9     s(anthro3a).1     s(anthro3a).2
##      1.119339e+00      1.000000e+00      2.738507e-08      6.645961e-10
##     s(anthro3a).3     s(anthro3a).4     s(anthro3a).5     s(anthro3a).6
##      3.972828e-09     -9.606393e-09      8.254896e-10     -1.102509e-08
##     s(anthro3a).7     s(anthro3a).8     s(anthro3a).9     s(anthro3c).1
##      8.717124e-11      7.292800e-08      1.000000e+00      9.935670e-01
##     s(anthro3c).2     s(anthro3c).3     s(anthro3c).4     s(anthro3c).5
##      1.017864e+00      9.098874e-01      6.369888e-01      6.302771e-01
##     s(anthro3c).6     s(anthro3c).7     s(anthro3c).8     s(anthro3c).9
##      1.839329e-01      3.318315e-01      1.337258e+00      1.000000e+00
```

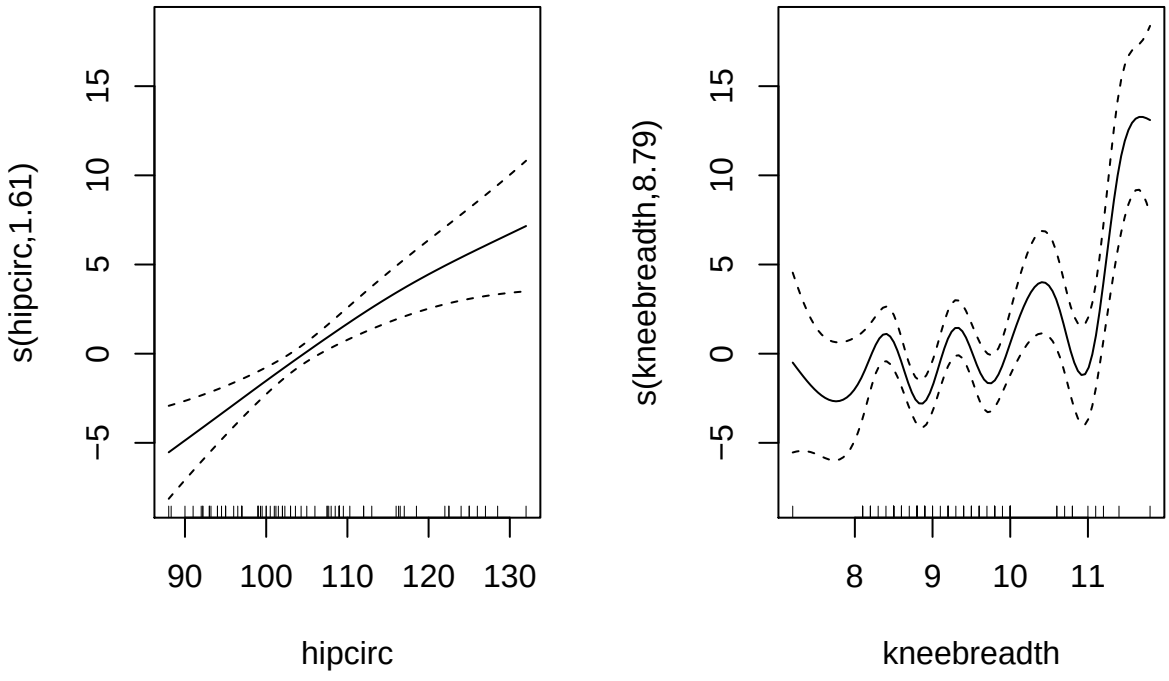
To find out whether the variables are informative or not, we look at the P-valueo f the variables. From the P-value, it seems waistcirc, hipcirc, kneebreadth, and anthro3a look very informative. These variables are highly significant to the model. From the graphs of gam, we could see that there is a linear relationship between the variables–age,waistcirc, hipcirc, elbowbreadth, anthro3a. Whereas kneebreadth, anthro2c has no linear relationship and hence needs smoothing. In the summary of the gam, some variables have an effective degree of freedom (edf) is 1 (one). The age, waistcirc, hipcirc, elbowbreadth and anthro3a has one as edf which means that these variables have a linear relationship (only). This same relationship is also shown in the graphs. Additionally, age, elbowbreadth, and anthro3c are not significant(0.05). Gam.check shows different plots, residual vs predicted model shows the random which no particular shape or pattern.The plot of response vs fitted values shows the positive upward linear relationship. In the histogram of residual, We can clearly see that the data is positively skewed.Summary shows that the GCV is 8.4354 which is moderate R- squared adjusted is 0.953 likewise AIC is 345.708 which is high.

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.

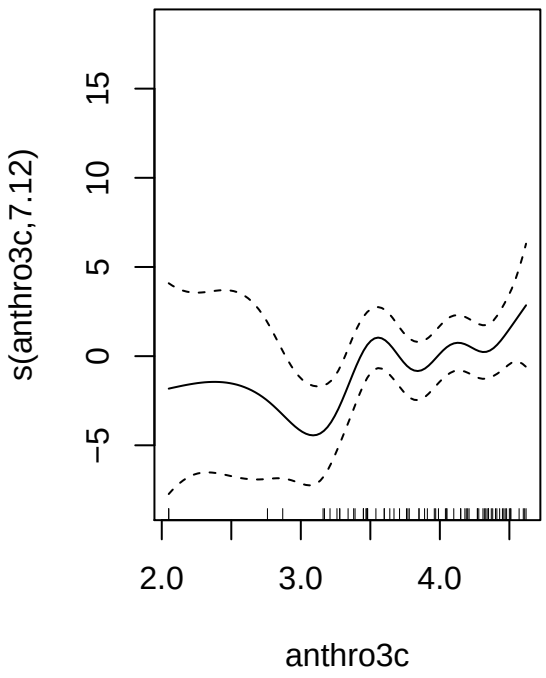
```
bodyfat_gam2 <- gam(DEXfat~ waistcirc + s(hipcirc) +
                    s(kneebreadth)+ anthro3a +
                    s(anthro3c), data = bodyfat)
```

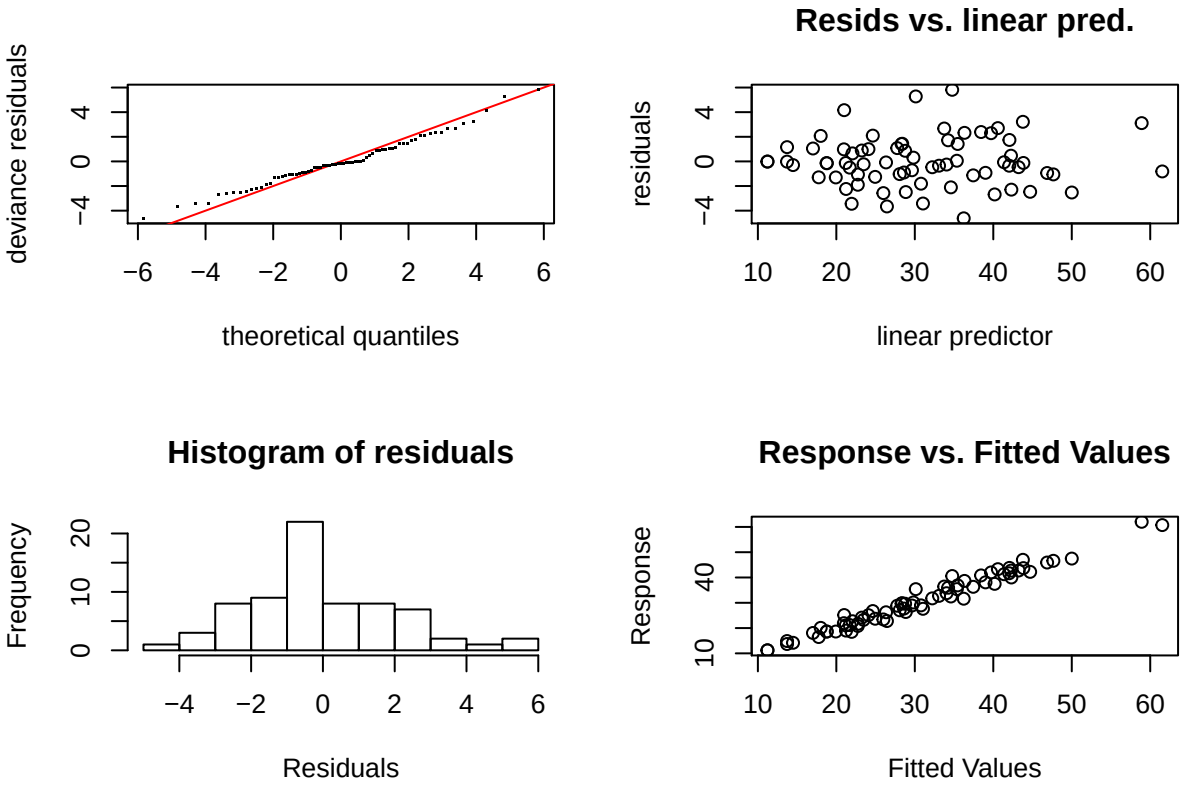
```
##
## Family: gaussian
## Link function: identity
```

```
##
## 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
```




```
## GCA for the bodyfat data: 7.946447
## AIC for the bodyfat data:  343.2562
```


s(anthro3c,7.12)
15
10
5
0
−5
2.0
3.0
4.0
anthro3c




```
##
## 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    0.48
## s(kneebreadth) 9.00 8.79    1.06    0.62
## s(anthro3c)    9.00 7.12    0.91    0.16
```

Here, insignificant variables like age, elbowbreadth are removed in this model. This change in a model doesn't add much value because GCV, R-sq(adjusted) remain unchanged. Also, the AIC of the second model decreased. From the graphs and the gam.check, we find out that hipcirc has a linear relationship. The only change we can see in the residual plot is in the histogram, the frequency of residual at the point zero to two decreased and becomes less than ten.

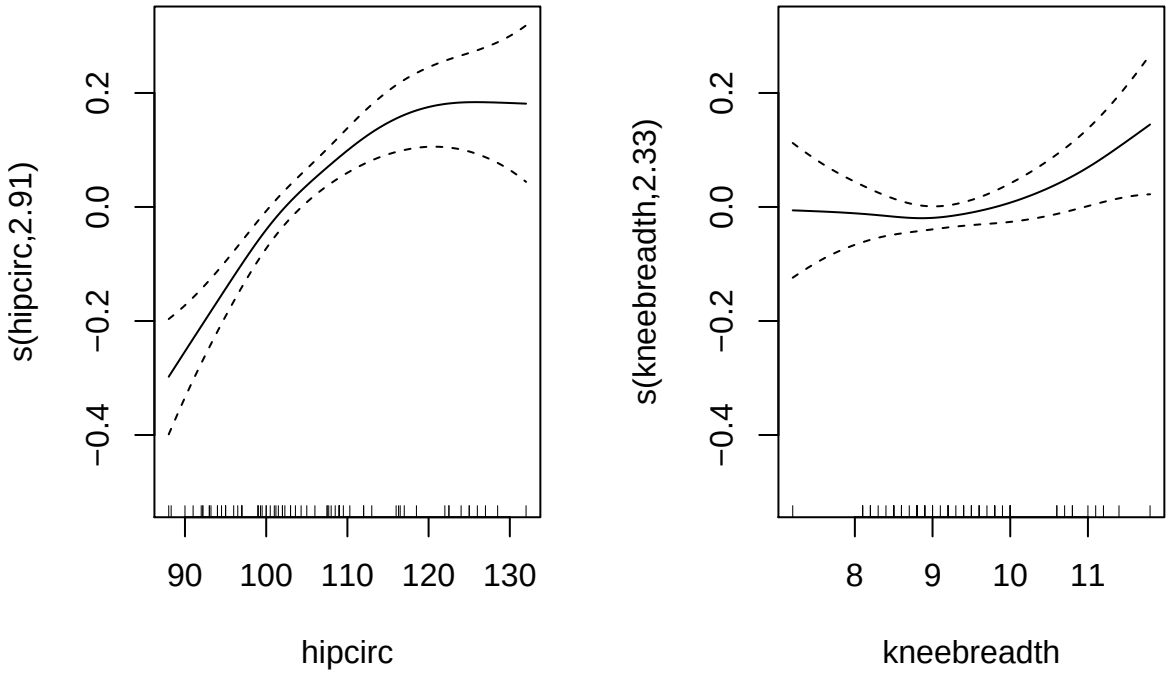
d) Again fit an additive model to the body fat data, but this time for a log-transformed response. Compare the three models, which one is more appropriate? (Hint: use AIC, GCV, residual plots, etc. to compare models).

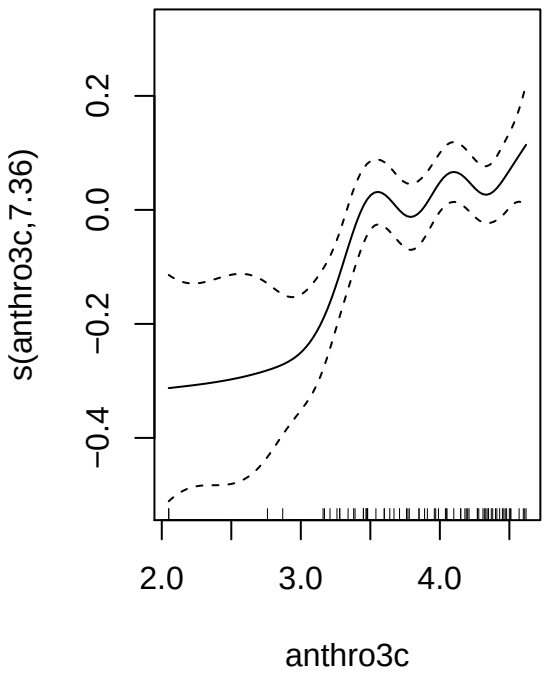
```
##
## Family: gaussian
## Link function: identity
```

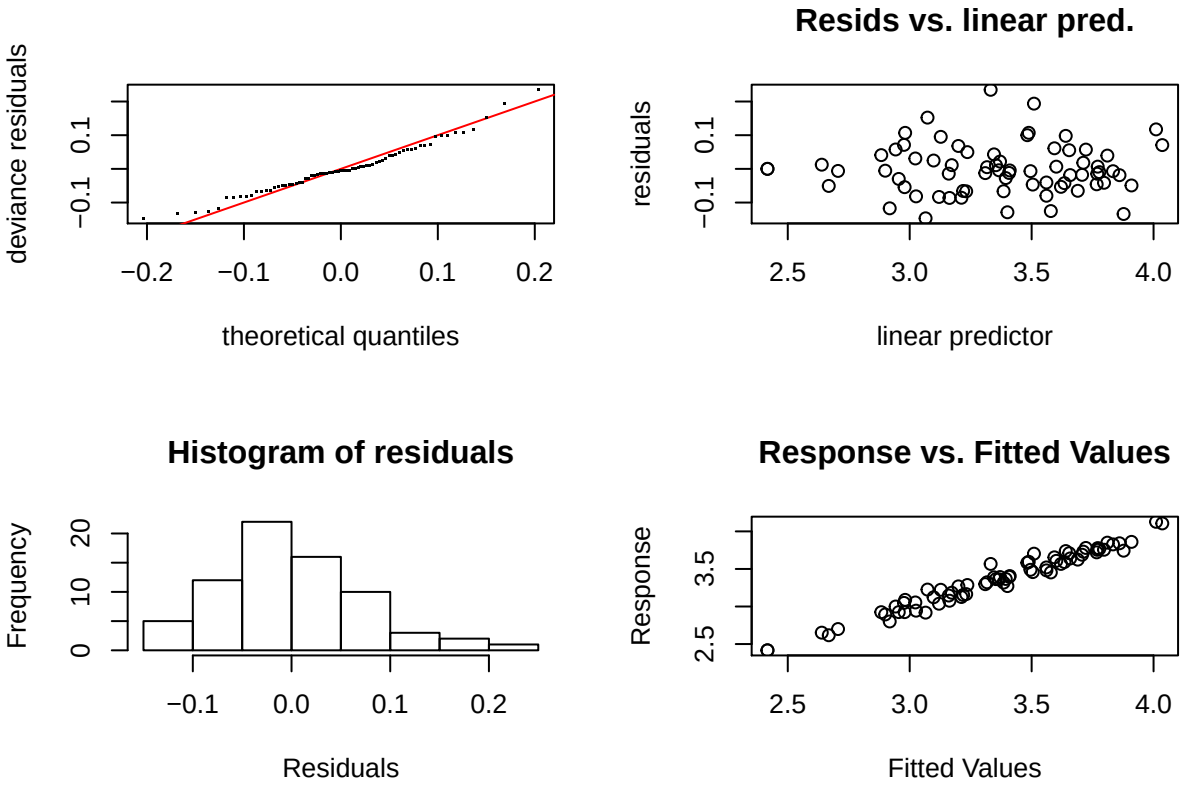
```
##
## Formula:
## logDex ~ waistcirc + s(hipcirc) + s(kneebreadth) + anthro3a +
##     s(anthro3c)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.139779   0.237083   9.025  1.8e-12 ***
## waistcirc   0.004418   0.001806   2.447 0.017610 *
## anthro3a    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 logDex;  -136.47

## GCV of logDex;  0.008813659
```




s(anthro3c,7.36)
−0.4
−0.2
0.0
0.2
2.0
3.0
4.0
anthro3c




```
##
## 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    0.12
## s(kneebreadth) 9.00 2.33    0.83    0.05 *
## s(anthro3c)    9.00 7.36    0.99    0.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

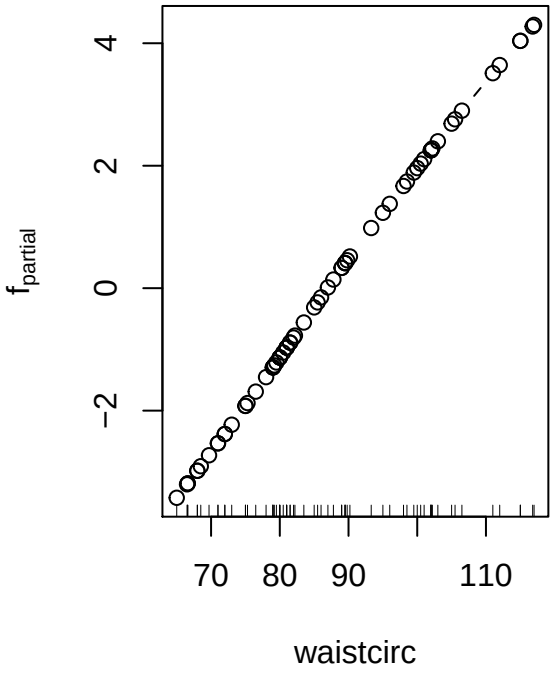
The log-transformed model shows the lesser GCV and negative AIC. However, R-squared has not changed in all three models. If I choose a model based on GCV and AIC, which is less for log-transformed and is the better model. However, if we consider R-squared (adjusted) that has not changed in any model. By looking at the residual plots, we can see that there is a decrease in scale in residual vs linear predictor. Likewise, the scale of the histogram's residual has changes(less).

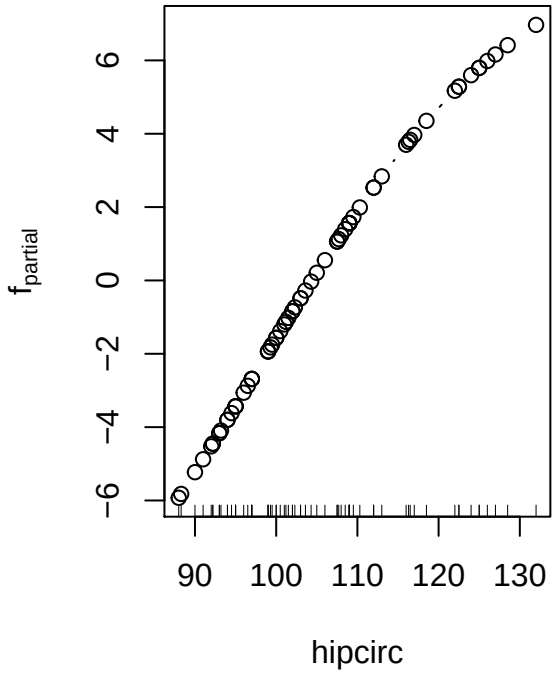
e) Run the code below to fit a generalised additive model that underwent AIC-based variable selection (fitted using the **gamboost()** function). What variable(s) was/were removed by using AIC?

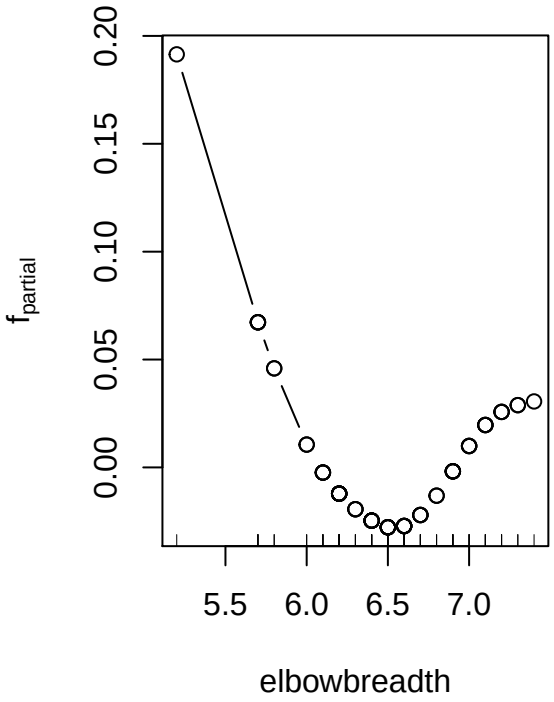
```
bodyfat_boost <- gamboost(DEXfat~., data = bodyfat)
bodyfat_aic <- AIC(bodyfat_boost)
```

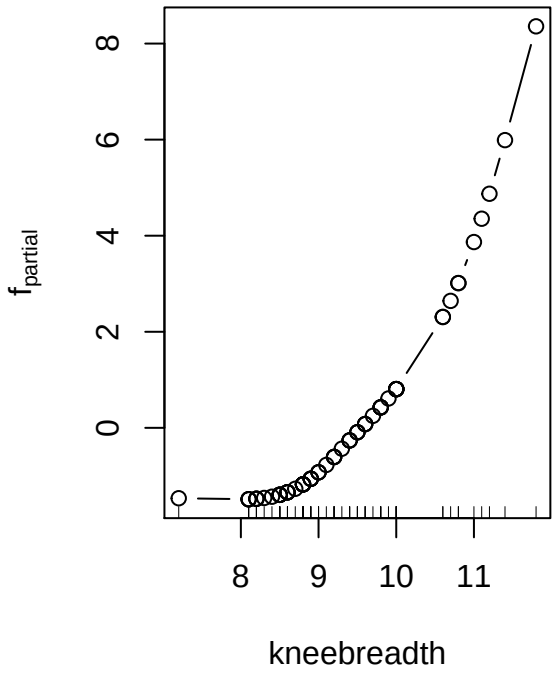
```
bf_gam <- bodyfat_boost[mstop(bodyfat_aic)]
```

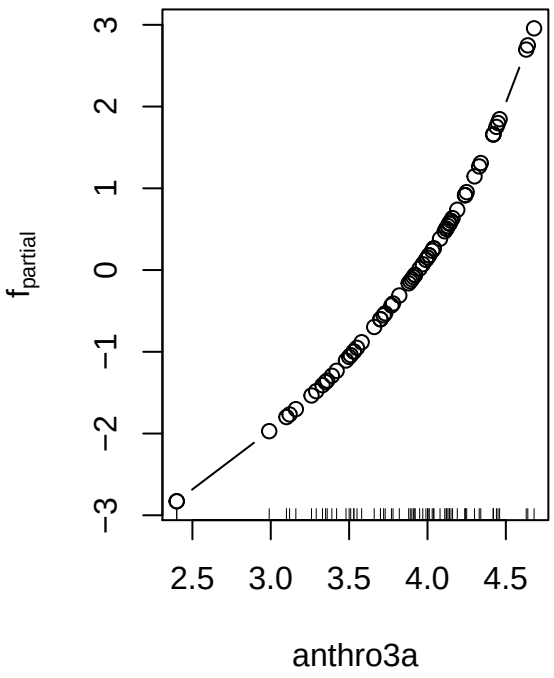
```
## [1] 3.268173
## Optimal number of boosting iterations: 51
## Degrees of freedom (for mstop = 51): 7.637287
```

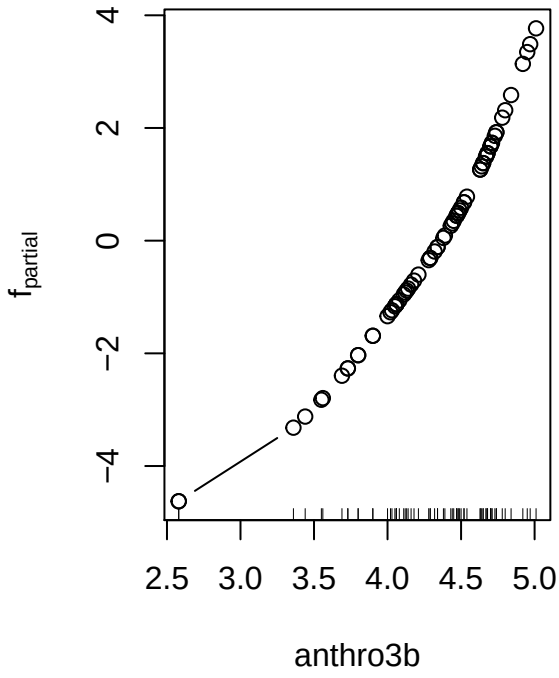


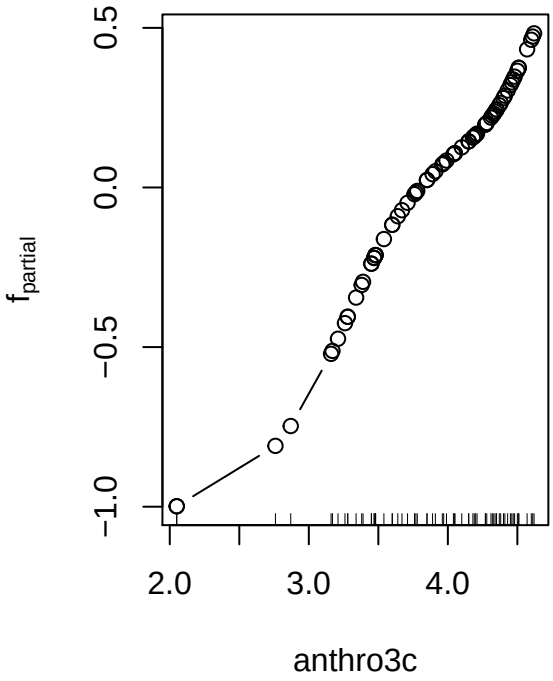


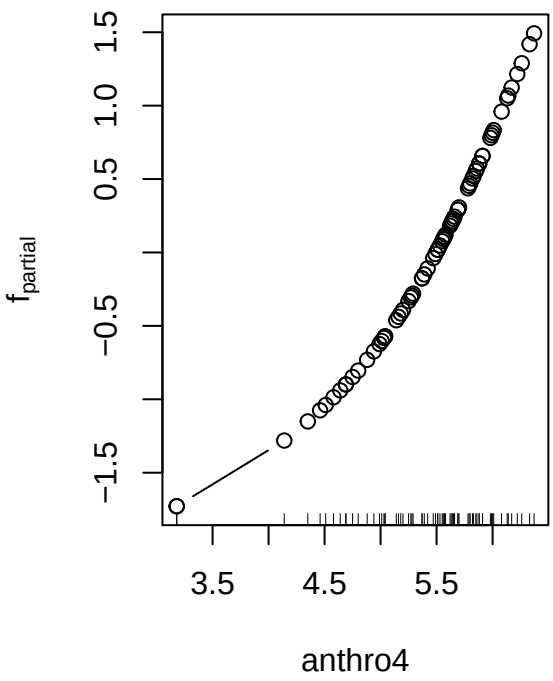

f_partial
0.00
0.05
0.10
0.15
0.20
5.5
6.0
6.5
7.0
elbowbreadth


f_partial
0
2
4
6
8
8
9
10
11
kneebreadth


f_partial
−3
−2
−1
0
1
2
3
2.5
3.0
3.5
4.0
4.5
anthro3a


f_partial
−4
−2
0
2
4
2.5
3.0
3.5
4.0
4.5
5.0
anthro3b






```
## extracting variable names
```

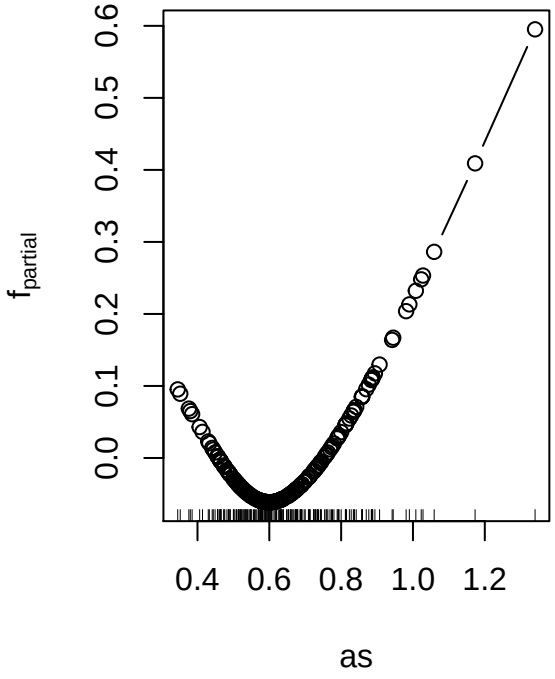
```
##    bbs(waistcirc, df = dfbase)      bbs(hipcirc, df = dfbase)
##                      "waistcirc"                        "hipcirc"
## bbs(elbowbreadth, df = dfbase)  bbs(kneebreadth, df = dfbase)
##                  "elbowbreadth"                   "kneebreadth"
##     bbs(anthro3a, df = dfbase)     bbs(anthro3b, df = dfbase)
##                       "anthro3a"                       "anthro3b"
##     bbs(anthro3c, df = dfbase)      bbs(anthro4, df = dfbase)
##                       "anthro3c"                       "anthro4"
```

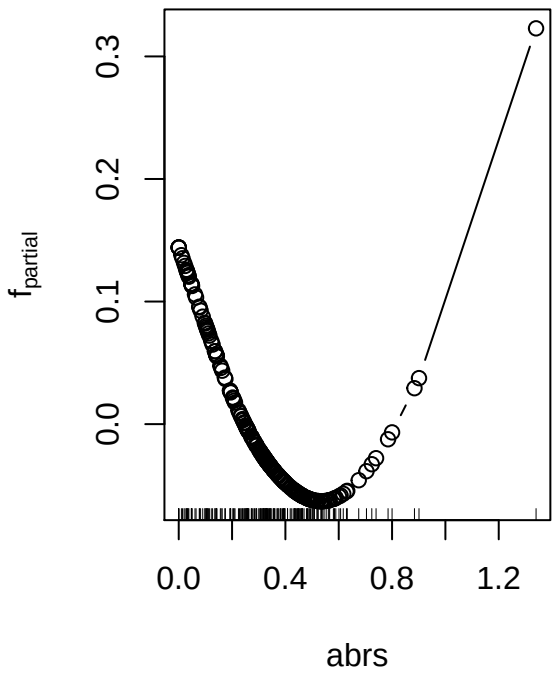
The variable Age was removed using AIC.

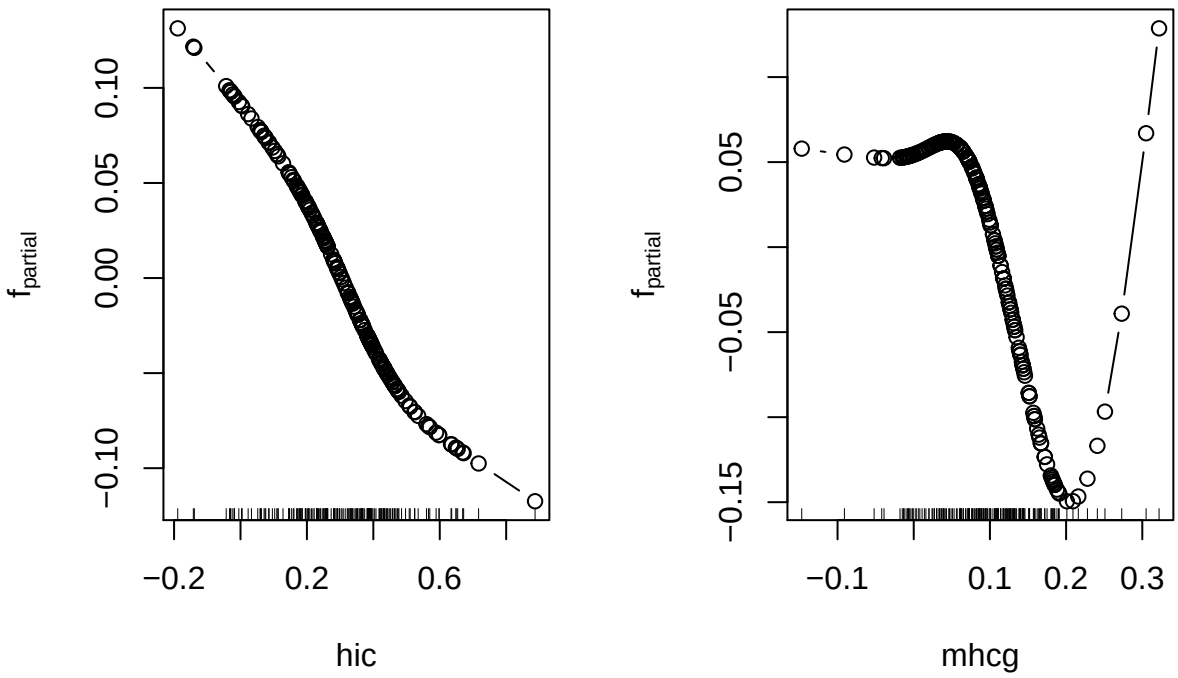
2. (Ex. 10.3 pg 208 in HSAUR, modified for clarity) Fit an additive model to the **glaucomaM** data from the **TH.data** library with *Class* as the response variable. Read the description of the dataset and the goals of the experiment. Which covariates should be in the model and what is their influence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

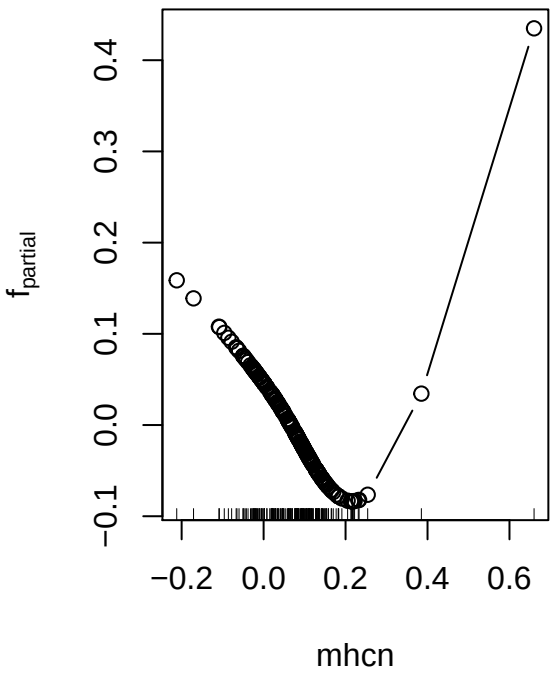
```
##
##   Model-based Boosting
##
## Call:
## gamboost(formula = Class ~ ., data = GlaucomaM, family = Binomial())
##
##
##   Negative Binomial Likelihood (logit link)
```

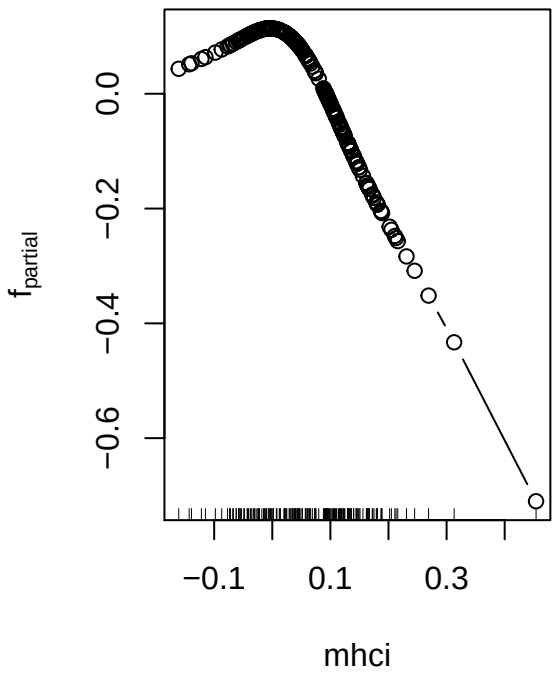
```
##
## Loss function: {
##      f <- pmin(abs(f), 36) * sign(f)
##      p <- exp(f)/(exp(f) + exp(-f))
##      y <- (y + 1)/2
##      -y * log(p) - (1 - y) * log(1 - p)
##  }
##
##
## Number of boosting iterations: mstop = 100
## Step size:  0.1
## Offset:  0
## Number of baselearners:  62
##
## Selection 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
```

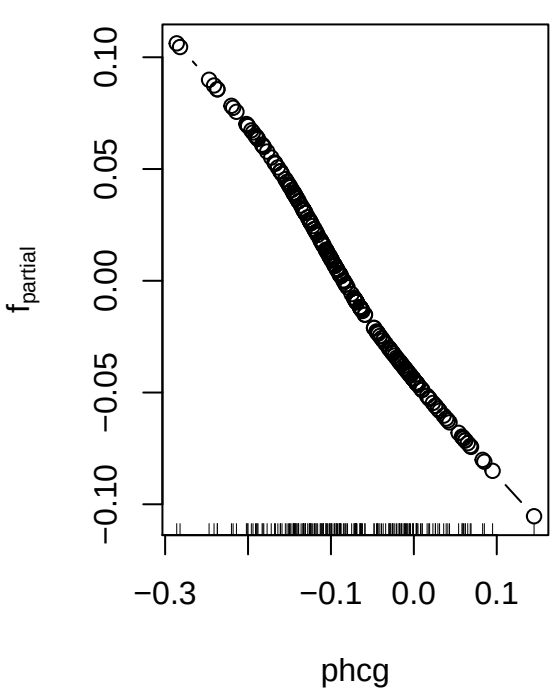

f_partial
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.4
0.6
0.8
1.0
1.2
as

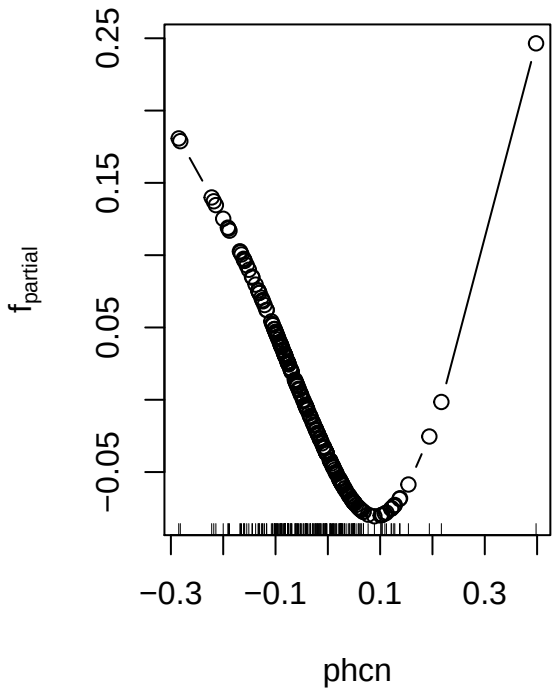

f_partial
0.0
0.1
0.2
0.3
0.0
0.4
0.8
1.2
abrs

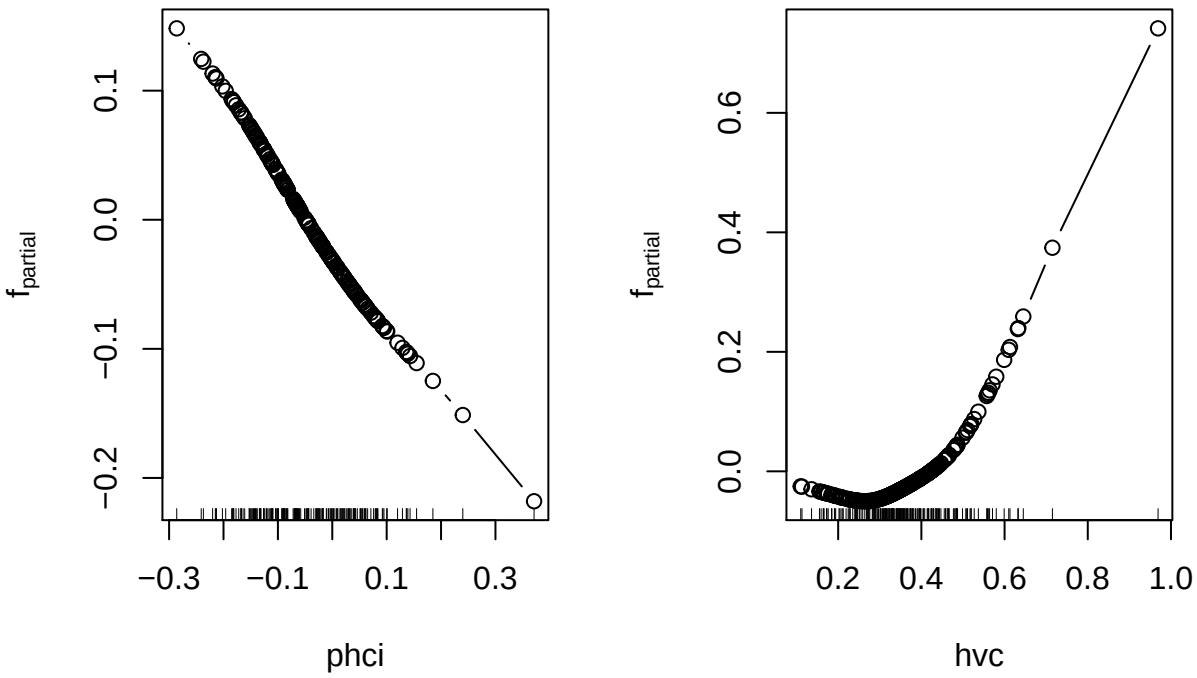

f_partial
0.10
0.05
0.00
−0.10
−0.2
0.2
0.6
hic
f_partial
0.05
−0.05
−0.15
−0.1
0.1
0.2
0.3
mhcg

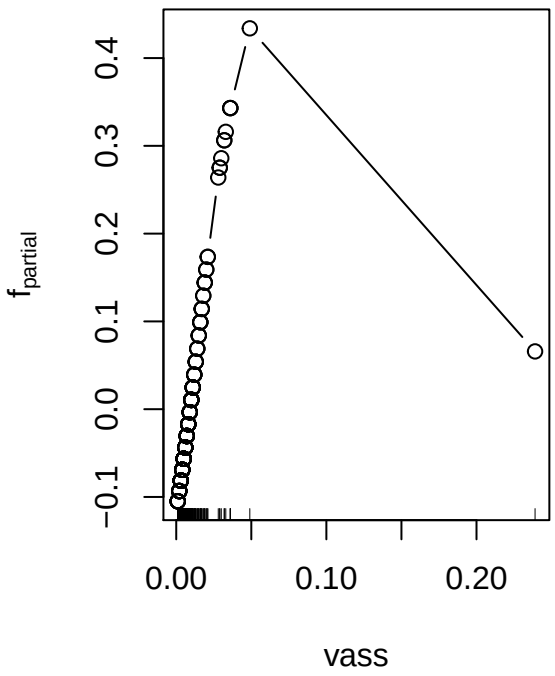

f_partial
−0.1
0.0
0.1
0.2
0.3
0.4
−0.2
0.0
0.2
0.4
0.6
mhcn

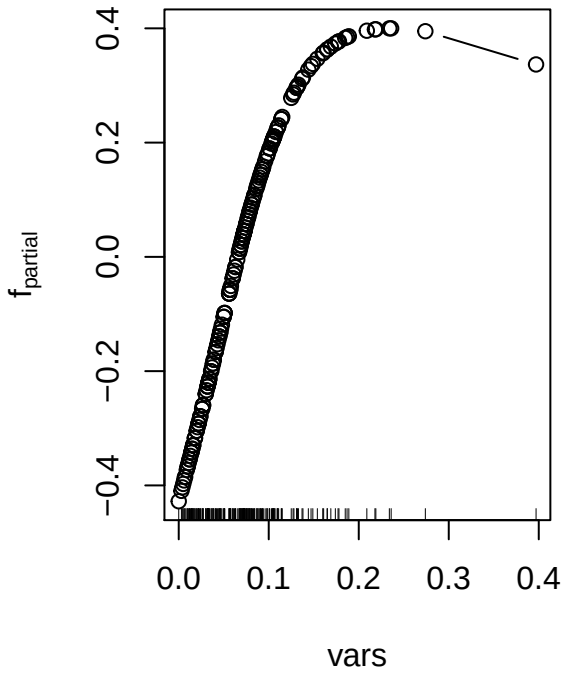

f_partial
−0.6
−0.4
−0.2
0.0
−0.1
0.1
0.3
mhci

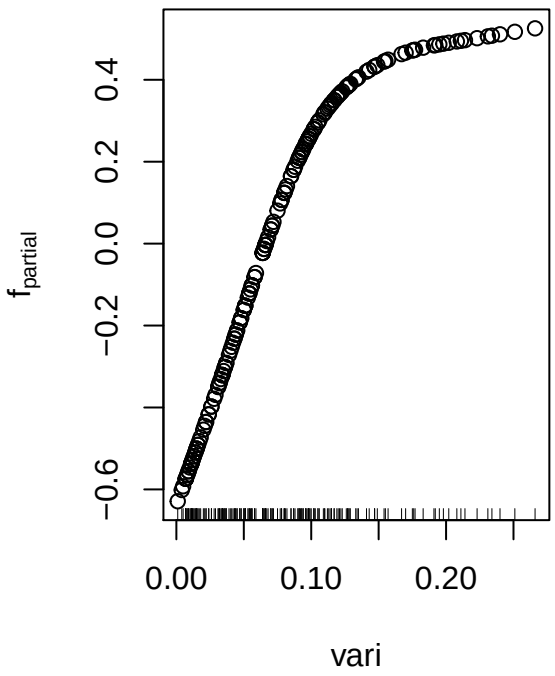

f_partial
0.10
0.05
0.00
−0.05
−0.10
−0.3
−0.1
0.0
0.1
phcg

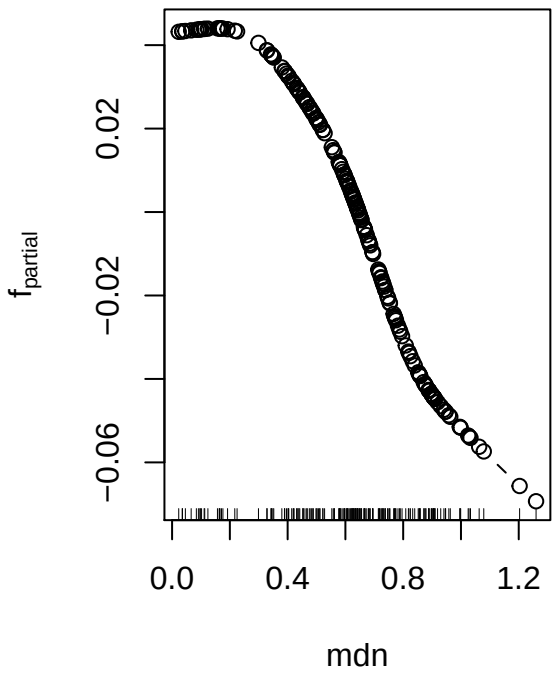

f_partial
0.25
0.15
0.05
−0.05
−0.3
−0.1
0.1
0.3
phcn

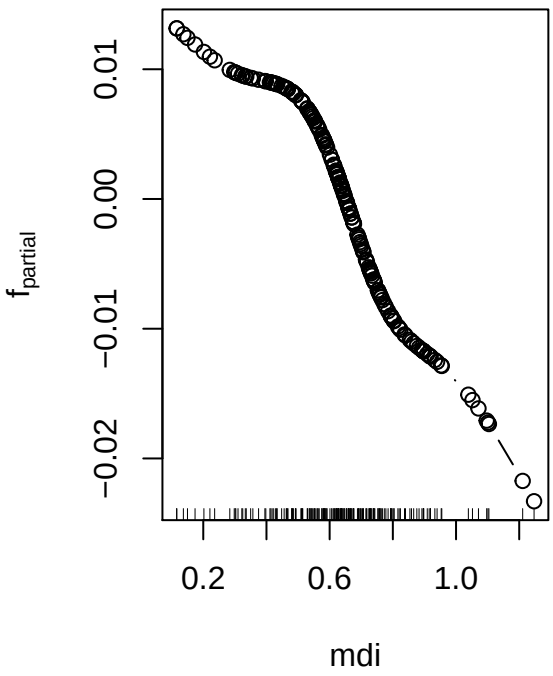

f_partial
0.1
0.0
−0.1
−0.2
−0.3
−0.1
0.1
0.3
phci
f_partial
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
hvc

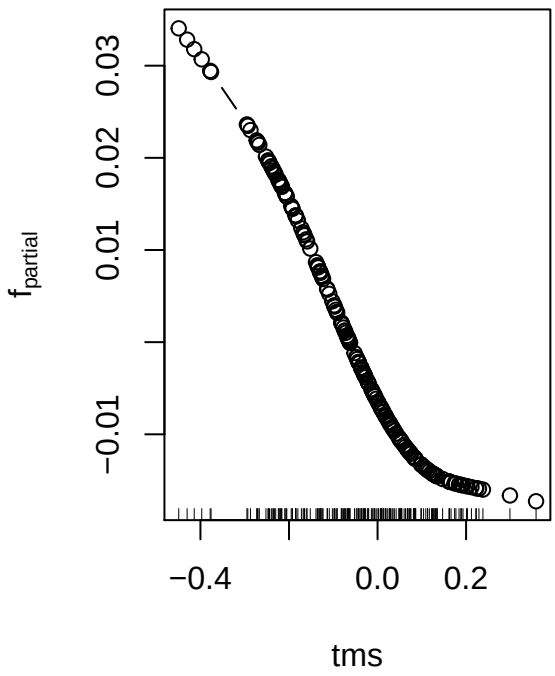

f_partial
−0.1
0.0
0.1
0.2
0.3
0.4
0.00
0.10
0.20
vass

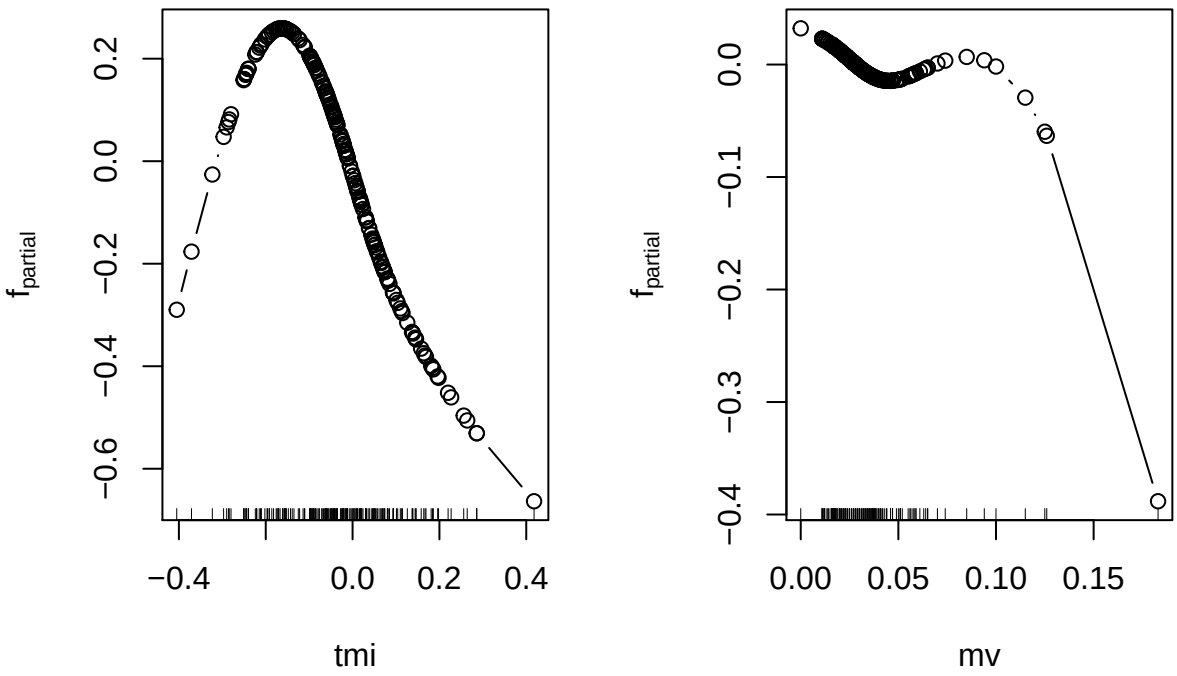

f_partial
−0.4
−0.2
0.0
0.2
0.4
0.0
0.1
0.2
0.3
0.4
vars


f_partial
−0.6
−0.2
0.0
0.2
0.4
0.00
0.10
0.20
vari


f_partial
−0.06
−0.02
0.02
0.0
0.4
0.8
1.2
mdn


f_partial
0.01
0.00
−0.01
−0.02
0.2
0.6
1.0
mdi


f_partial
0.03
0.02
0.01
−0.01
−0.4
0.0
0.2
tms




```
##  [1] "as"   "abrs" "hic"  "mhcg" "mhcn" "mhci" "phcg" "phcn" "phci" "hvc"
## [11] "vass" "vars" "vari" "mdn"  "mdi"  "tms"  "tmi"  "mv"
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Class ~ s(abrs) + s(as) + s(hic) + s(hvc) + s(mdi) + s(mdn) +
##     s(mhcg) + s(mhci) + s(mhcn) + s(mv) + phcg + phci + s(phcn) +
##     s(tmi) + s(tms) + s(vari) + s(vars) + s(vass)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    15.65    1856.10   0.008    0.993
## phcg          602.17   11941.01   0.050    0.960
## phci         -295.39   11159.97  -0.026    0.979
##
## Approximate significance of smooth terms:
##           edf Ref.df Chi.sq p-value
## s(abrs) 1.000  1.000  0.001   0.972
## s(as)   1.393  1.429  0.006   0.987
## s(hic)  1.000  1.000  0.002   0.966
## s(hvc)  5.194  5.266  0.012   1.000
## s(mdi)  1.000  1.000  0.001   0.972
## s(mdn)  1.000  1.000  0.000   0.989
## s(mhcg) 1.000  1.000  0.001   0.974
```

```
## s(mhci) 2.738  2.815  0.002   1.000
## s(mhcn) 1.328  1.356  0.000   0.996
## s(mv)   1.000  1.000  0.006   0.938
## s(phcn) 3.975  4.024  0.003   1.000
## s(tmi)  2.456  2.517  0.010   0.999
## s(tms)  1.000  1.000  0.000   0.999
## s(vari) 3.592  3.680  0.010   1.000
## s(vars) 1.000  1.000  0.001   0.970
## s(vass) 1.000  1.000  0.014   0.905
##
## R-sq.(adj) =      1   Deviance explained =  100%
## UBRE = -0.66687  Scale est. = 1         n = 196
```

The variables 'as" , "abrs", "hic" , "mhcg", "mhcn", "mhci", "phcg", "phcn", "phci", "hvc", "vass", "vars", "vari", "mdn", "mdi", "tms","tmi" should be in the model and they has their influence on the probability of suffering from glaucoma. From the graph of glucoma data it is clear that phcg and phci has the liner relationship.This is why I removes s in the gam model. R squared of this over fitted is 1 and Deviance of this model is 100% which means that the model is extremly over fitted.