Homework 6

Alex Soupir

October 20, 2019

Packages: HSAUR3, mgcv, GGally, mboost, rpart, wordcloud, ggplot2, TH.data, tidyverse, gamair

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

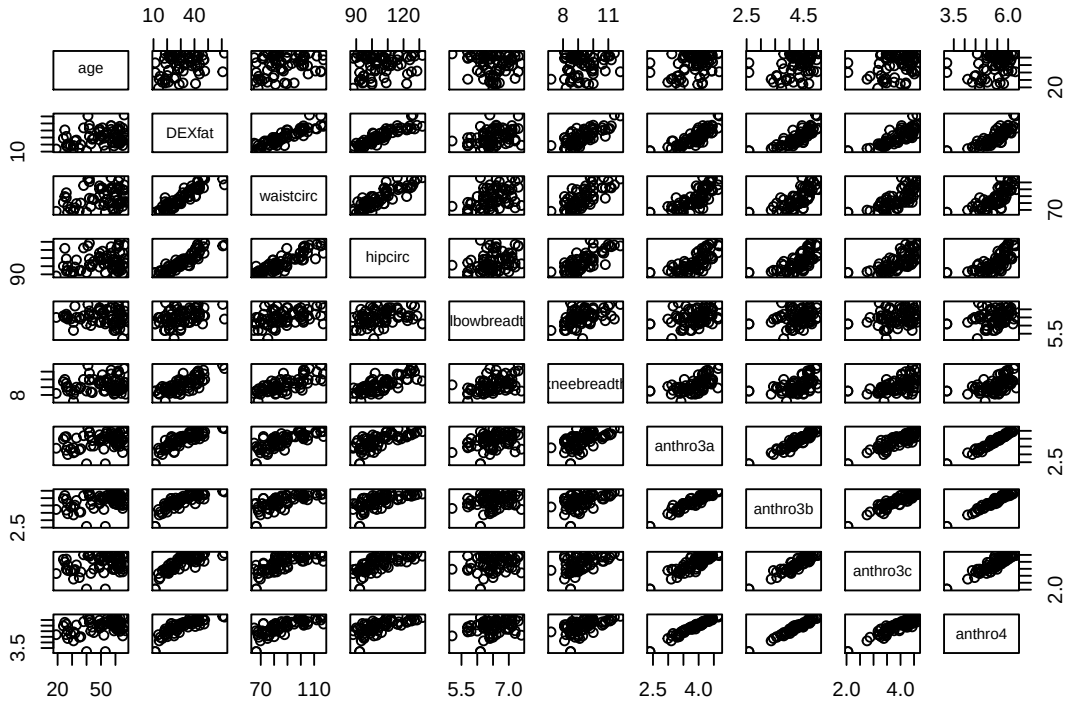
For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

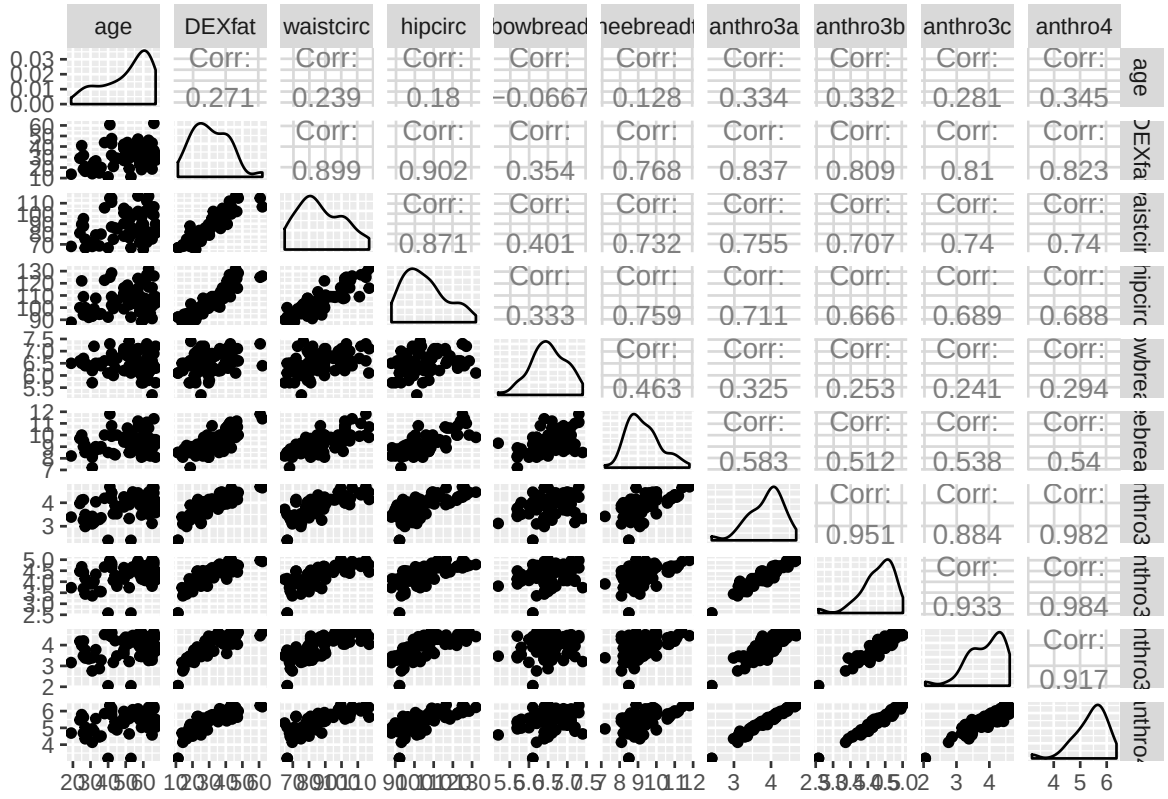
You do not need to include the above statements.

Please do the following problems from the text book R Handbook and stated.

1. Consider the body fat data introduced in Chapter 9 (**bodyfat** data from **TH.data** package).

    a) Explore the data graphically. What variables do you think need to be included for predicting bodyfat? (Hint: Are there correlated predictors).





As with last weeks assignment, there are a few variables that appear to be highly correlated to each other. The suggestion was to remove these.

```
## [1] "age"          "waistcirc"    "hipcirc"      "elbowbreadth"
## [5] "kneebreadth"  "anthro3a"     "anthro3c"     "DEXfat"
```

Since we want to have both waistcirc, anthro3c, and hipcirc in part b, the threshold for correlation elimination was set to 0.94 to remove those that are highly correlated while maintaining these 3 features (anthro3c has a correlation of 0.933 with anthro3b, I believe, so I set the threshold slightly higher). The variables that I think need to be included for predicting bodyfat given the above is *age*, *waistcirc*, *hipcirc*, *elbowbreadth*, *kneebreadth*, *anthro3a*, and *DEXfat*. This means that Antro

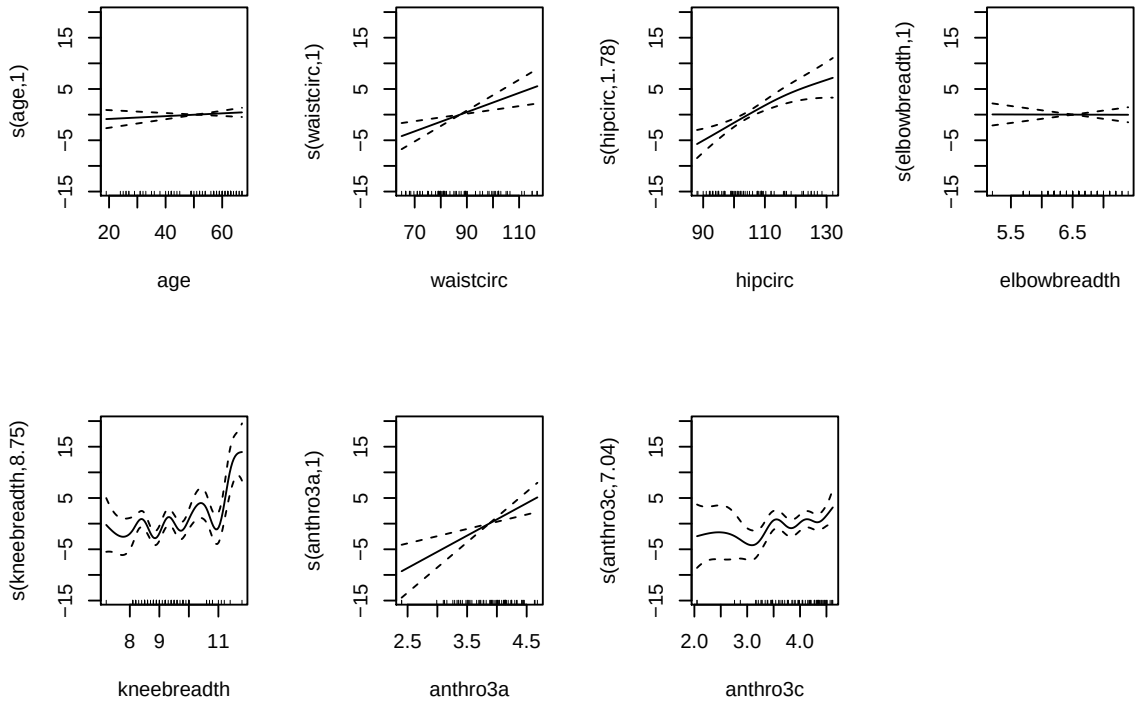
b) Fit a generalised additive model assuming normal errors using the following code.

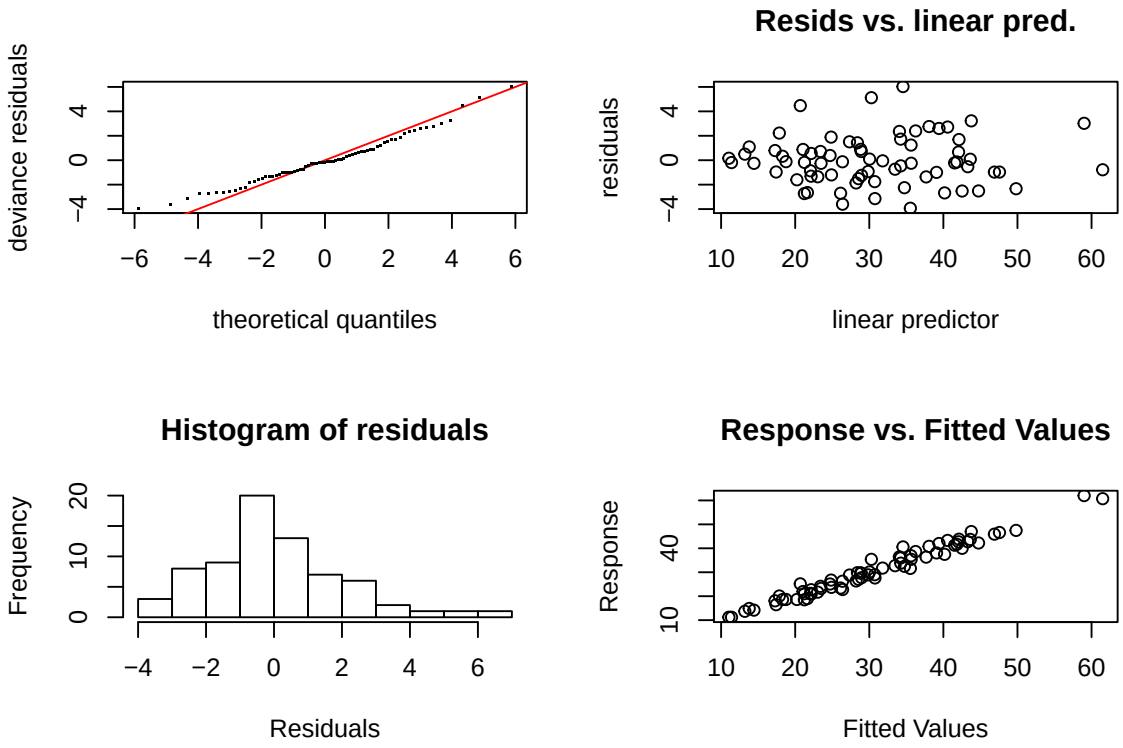
```
bodyfat_gam <- gam(DEXfat~ s(age) + s(waistcirc) + s(hipcirc) +
         s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) +
         s(anthro3c), data = bodyfat)
```

- Assess the **summary()** and **plot()** of the model (don't need GGPLOT). Are all covariates informative? Should all covariates be smoothed or should some be included as a linear effect?
- Report GCV, AIC, adj-R^2, and total model degrees of freedom.
- Use **gam.check()** function to look at the diagnostic plot. Does it appear that the normality assumption is violated?
- 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
##
## Model AIC:
```

```
## [1] 345.708
## Estimated Degrees of Freedom:
## [1] 22.57091
```






```
##
## 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   0.045 *
## s(waistcirc)    9.00 1.00    0.94   0.320
## s(hipcirc)      9.00 1.78    1.02   0.555
## s(elbowbreadth) 9.00 1.00    0.81   0.050 *
## s(kneebreadth)  9.00 8.75    1.08   0.665
## s(anthro3a)     9.00 1.00    1.09   0.745
## s(anthro3c)     9.00 7.04    0.89   0.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

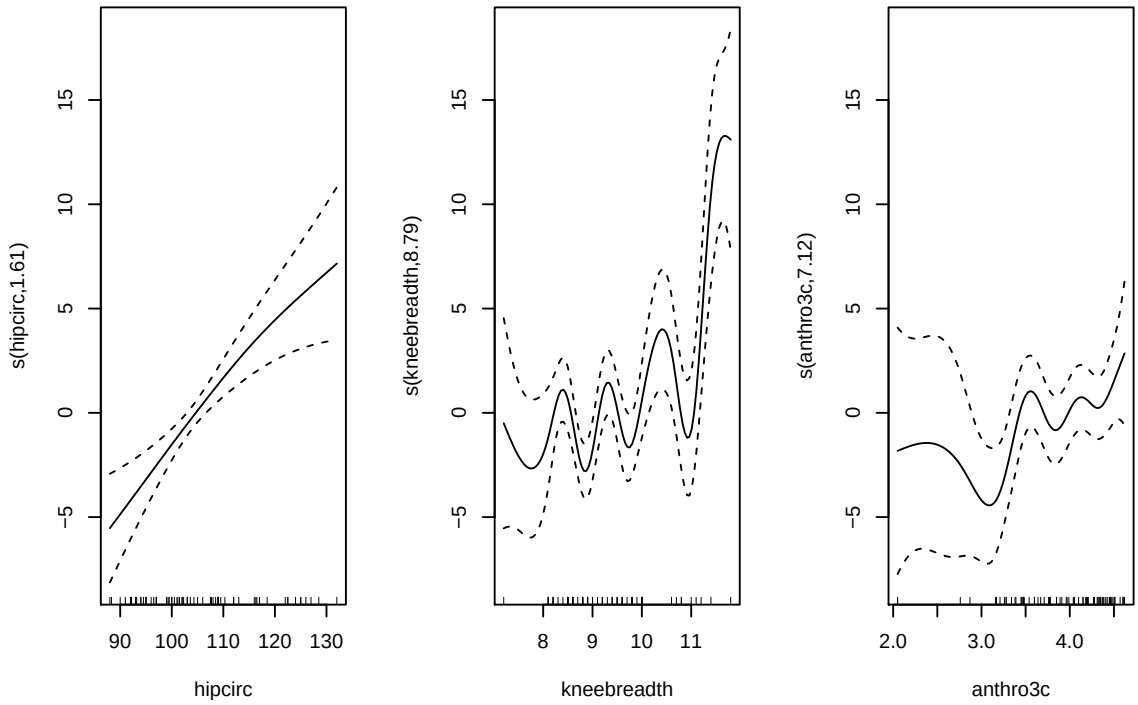
GCV = 8.44, AIC = 345.71, adj-R^2 = 0.95, model degrees of freedom = 22.57 (including 1 df for intercept as per stackexchange, or 21.57 not including 1 for the intercept).Variables *age*, *elbowbreadth*, and *anthro3c* are not significant at p = 0.05. Variables *age*, *waistcirc*, *elbowbreadth*, and *anthro3a* have an estimated degrees of freedom of 1 while *kneebreadth* has an edf of 8.754. Comparing these values to the plot, the higher the degrees of freedom the less linear the smoothed {s()} variable is, so *age*, *waistcirc*, *elbowbreadth*, and *anthro3a* could be included as a linear effect. The models generalized cross-validation

score (GCV) is 8.44 and AIC is 345.71 on 22.57 degrees of freedom. The GCV and AIC most likely could be improved through feature selection than smoothing those features in the generalized additive model.

c) Now remove insignificant variables and remove smoothing for some variables. Report the summary, plot, GCV, AIC, adj-R^2.

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

```
## Given 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 0.000103 ***
## 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
```



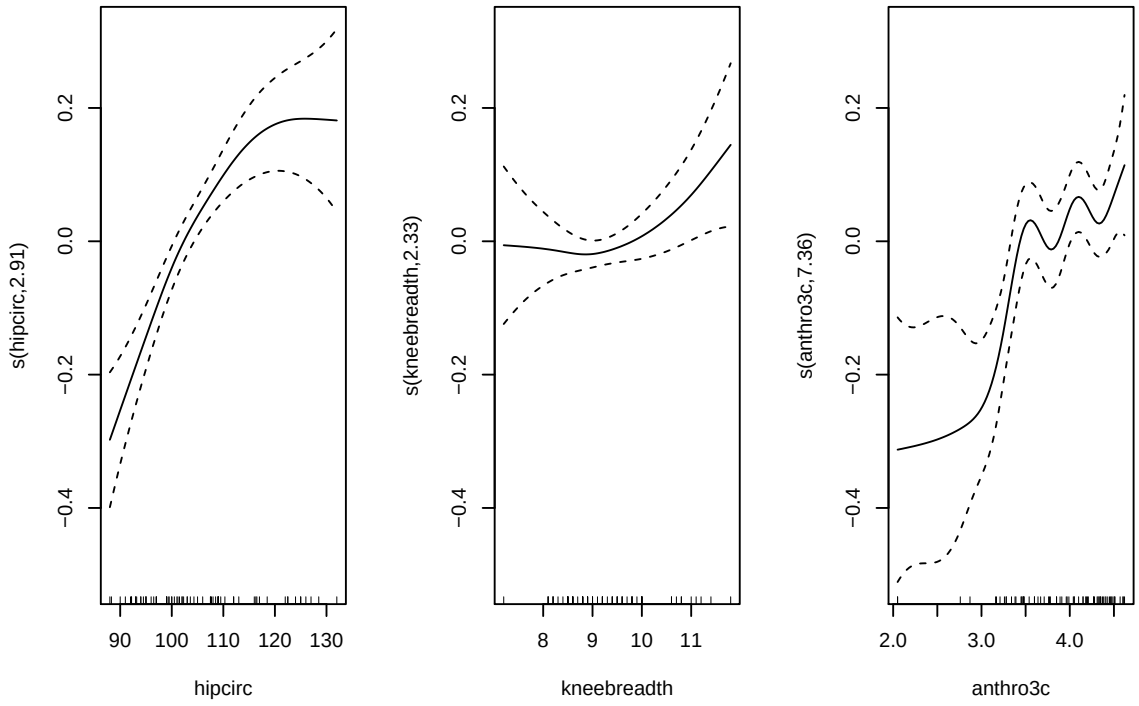

```
##
## Given Model AIC:
## [1] 343.2562
## Estimated Degrees of Freedom:
## [1] 20.52001
```

GCV = 7.95, AIC = 343.26, adj-R^2 = 0.95. All variable terms were significant at p = 0.05 and the intercept was significant at p = 0.1 level. The generalized cross-validation score was better than that of the previous model at 7.95 instead of 8.44. The AIC of the model with only including the significant variables from the previous model decreased slightly which indicates that the new model performs better. However, the adj-R^2 increased by 0.001

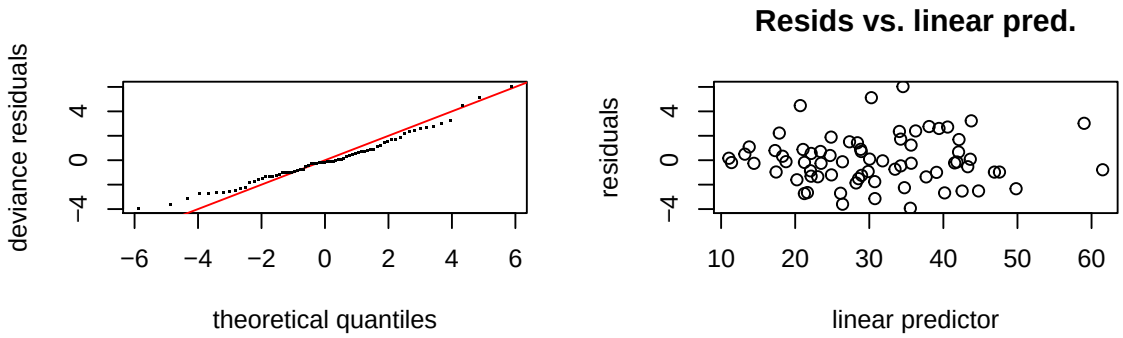
d) Again fit an additive model to the body fat data, but this time for a log-transformed response. Compare the three models, which one is more appropriate? (Hint: use Adj-R^2, residual plots, etc. to compare models).

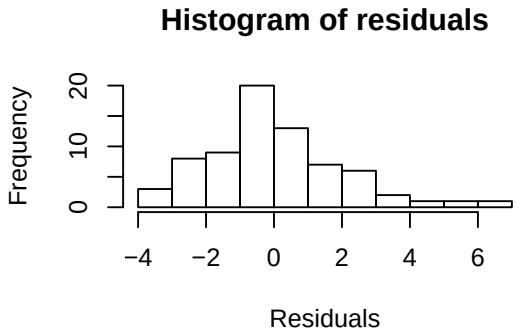
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## log(DEXfat) ~ waistcirc + s(hipcirc) + s(kneebreadth) + anthro3a +
##     s(anthro3c)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.139779   0.237083   9.025  1.8e-12 ***
```

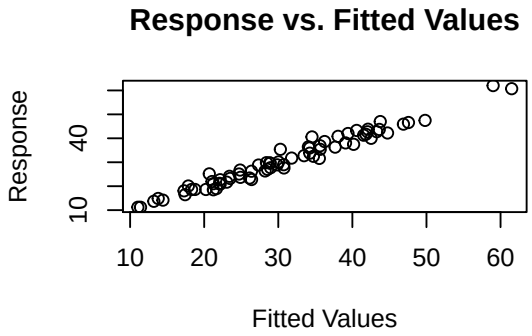
```
## waistcirc   0.004418   0.001806   2.447 0.017610 *
## anthro3a    0.215488   0.054600   3.947 0.000226 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                  edf Ref.df      F  p-value
## s(hipcirc)     2.909  3.616 11.828  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
```



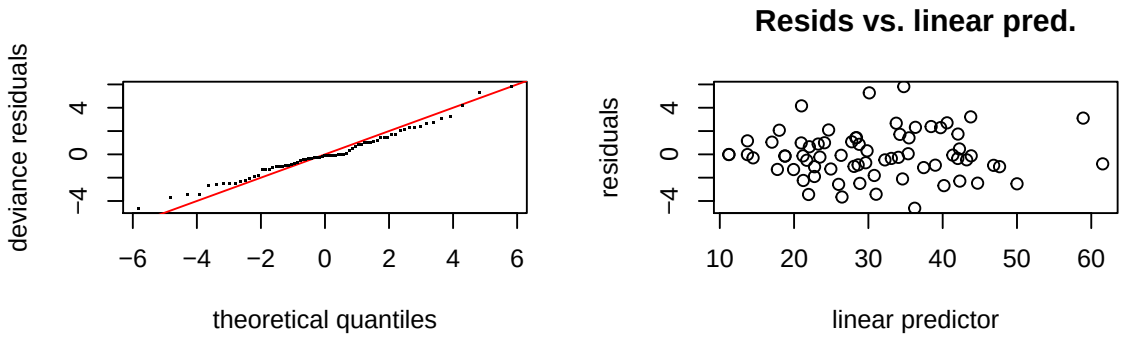

```
##
## Log transformed response model AIC:
## [1] -136.47
## Estimated Degrees of Freedom:
## [1] 15.59274
##
## Part B Model:
```

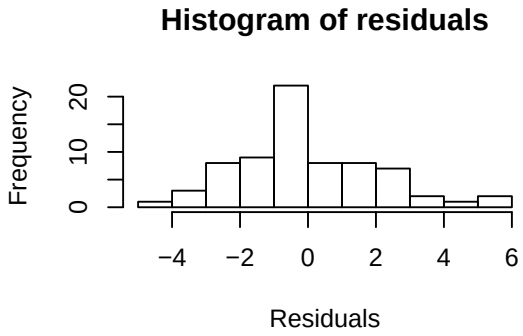


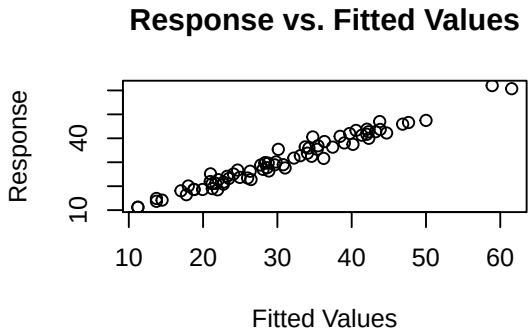




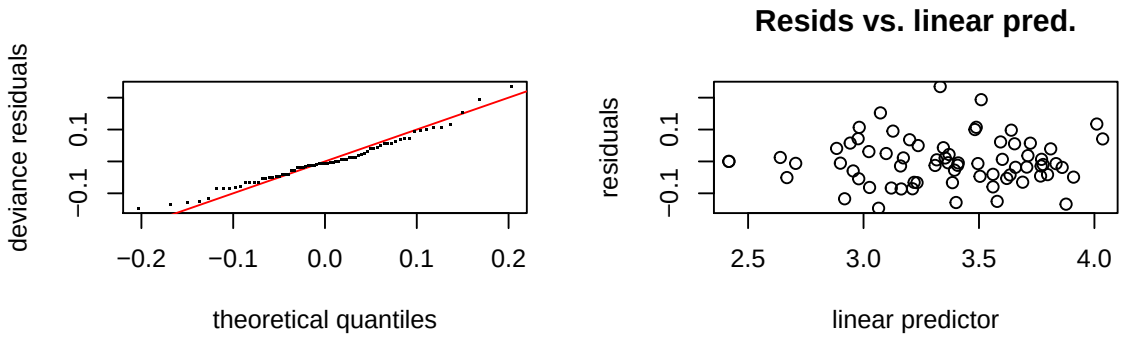

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 41 iterations.
## The RMS GCV score gradient at convergence was 2.767255e-07 .
## The Hessian was positive definite.
## Model rank =  64 / 64
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## 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   0.320
## s(hipcirc)      9.00 1.78    1.02   0.555
## s(elbowbreadth) 9.00 1.00    0.81   0.050 *
## s(kneebreadth)  9.00 8.75    1.08   0.665
## s(anthro3a)     9.00 1.00    1.09   0.745
## s(anthro3c)     9.00 7.04    0.89   0.195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Part C Model:
```

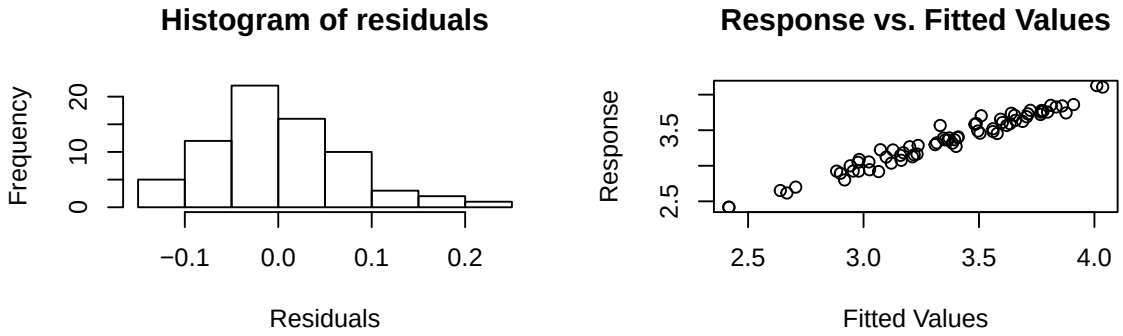







```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 24 iterations.
## The RMS GCV score gradient at convergence was 0.0001386163 .
## The Hessian was positive definite.
## Model rank =  30 / 30
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## 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    0.52
## s(kneebreadth) 9.00 8.79    1.06    0.66
## s(anthro3c)    9.00 7.12    0.91    0.20

##
## Part D Model:
```






```
##
## 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   0.045 *
## s(anthro3c)    9.00 7.36    0.99   0.460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The log transformed response variable decreases the generalized cross-validation score and the AIC of the model. The log transformation of the response variable produces a GCV of 0.009 AIC of -136.5 with an estimated 15.6 degrees of freedom. The smoothing of the kneebreadth variable isn't significant at p=0.05 now too.

e) Fit a generalised additive model that underwent AIC-based variable selection (fitted using function **gamboost()** function). What variable was removed by using AIC?

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

The variables that were used through the selection are *waistcirc*, *hipcirc*, *elbowbreadth*, *kneebreadth*, *anthro3a*, *anthro3b*, *anthro3c*, and *anthro4*. This shows that the variable that was eliminated using AIC was the age variable.

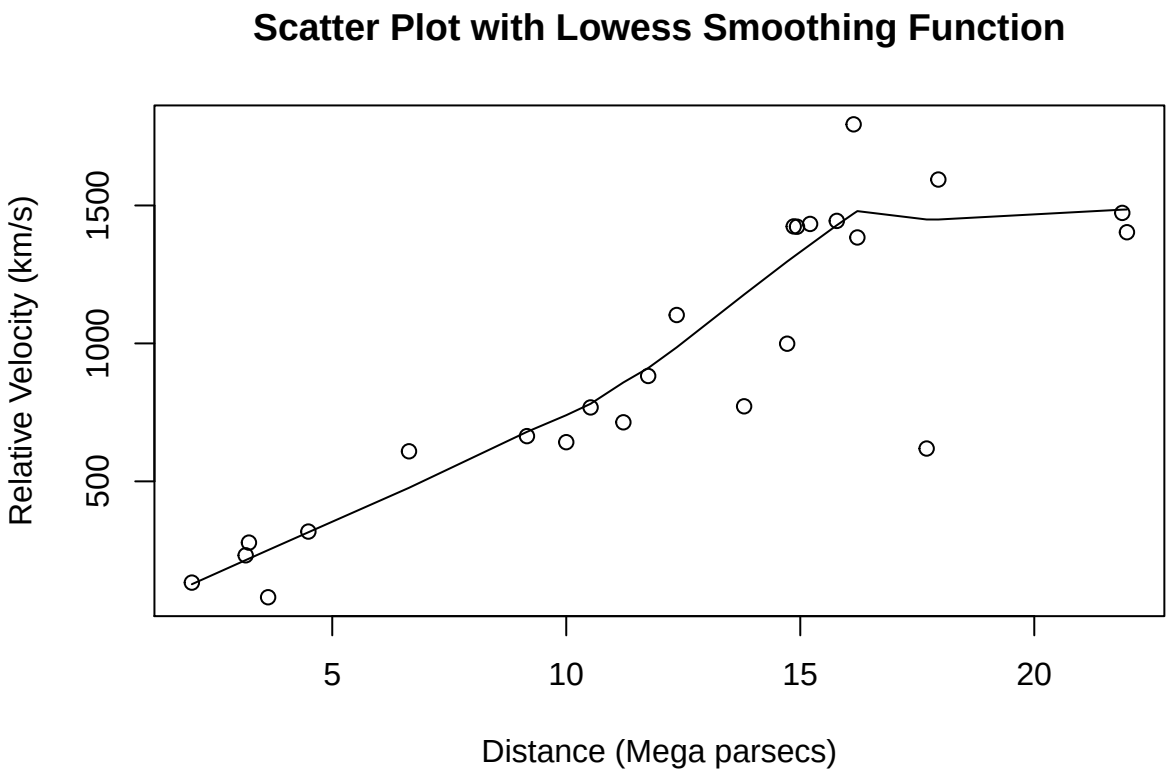
2. Fit a logistic additive model to the glaucoma data. (Here use family = "binomial"). Which covariates should enter the model and how is their influence on the probability of suffering from glaucoma? (Hint: since there are many covariates, use **gamboost()** to fit the GAM model.)

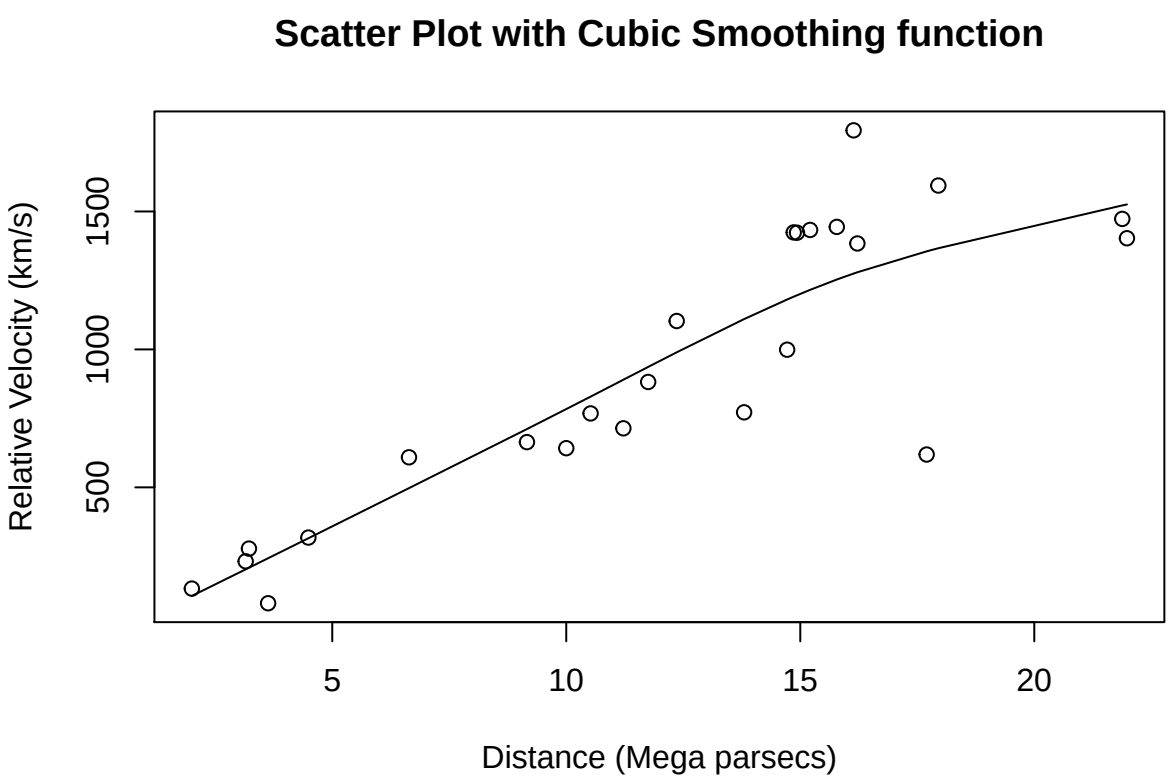
```
##
##   Model-based Boosting
##
## Call:
## gamboost(formula = Class ~ ., data = glau, family = Binomial())
##
##
##   Negative Binomial Likelihood (logit link)
##
## Loss function: {
##      f <- pmin(abs(f), 36) * sign(f)
##      p <- exp(f)/(exp(f) + exp(-f))
##      y <- (y + 1)/2
##      -y * log(p) - (1 - y) * log(1 - p)
##  }
##
##
## Number of boosting iterations: mstop = 100
## Step size:  0.1
## Offset:  0
```

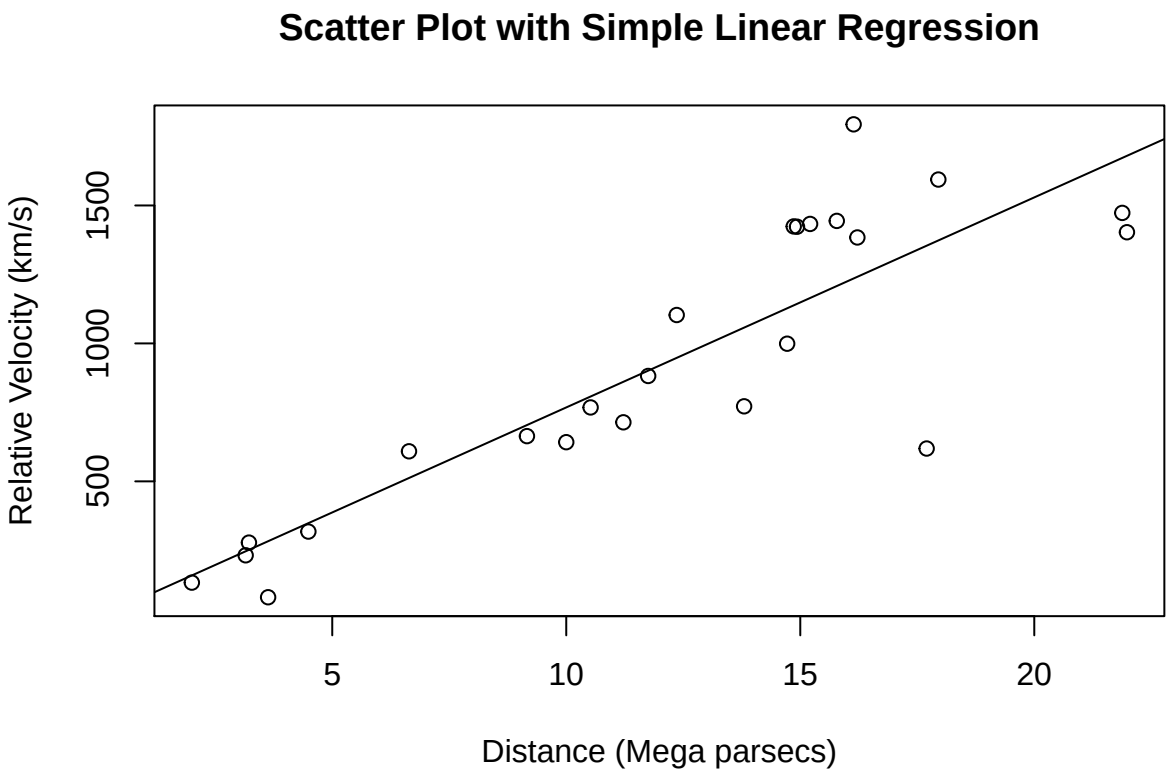
```
## Number of baselearners:  62
##
## Selection frequencies:
##  bbs(tmi, df = dfbase) bbs(mhcg, df = dfbase) bbs(vars, df = dfbase)
##                   0.17                   0.11                   0.11
## bbs(mhci, df = dfbase)  bbs(hvc, df = dfbase) bbs(vass, df = dfbase)
##                   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

##
## Means Squared Error of Predictions on Glaucoma DF using Boost Model:

## [1] 0.1020408
```

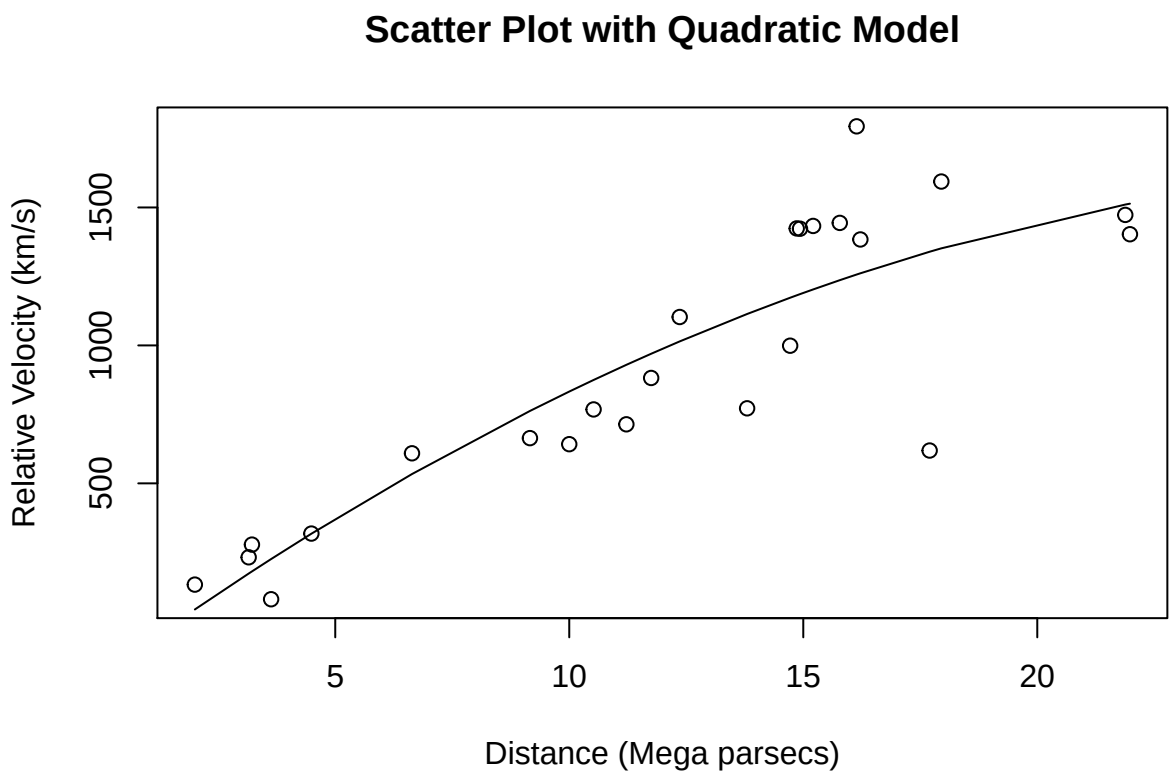
The variable selected using *gamboost()* are *tmi*, *mhcg*, *vars*, *mhci*, *hvc*, *vass*, *as*, *vari*, *mv*, *abrs*, *mhcn*, *phcn*, *mdn*, *phci*, *hic*, *phcg*, *mdi*, and *tms*. Using these variables, the error of the model is lower than that of the all models created last week for the glaucoma data set except the adaptive boosting on the whole data which perfectly fit the model. A train test split of 80/20 resulted in an error (MSE was the same value as the error) of 0.195, whereas gam boosting on the data produced an MSE of 0.10. This means that the variables are better able to predict whether a patient has glaucoma or not than k-fold cross validation (k-10, MSE=0.27) and glm (MSE=0.48.

3. Investigate the use of different types of scatterplot smoothers on the Hubble data from Chapter 6. (Hint: follow the example on men1500m data scattersmoothers page 199 of Handbook).


Scatter Plot with Lowess Smoothing Function
Relative Velocity (km/s)
500
1000
1500
5
10
15
20
Distance (Mega parsecs)


Scatter Plot with Cubic Smoothing function
Relative Velocity (km/s)
500
1000
1500
5
10
15
20
Distance (Mega parsecs)


Scatter Plot with Simple Linear Regression
Relative Velocity (km/s)
500
1000
1500
5
10
15
20
Distance (Mega parsecs)



**The cubic model

Resources Used:

- stat.ethz.ch
- rdocumentation.org
- stackexchange.com
- Homework 5