

Homework 6

Alex Soupir

October 20, 2019

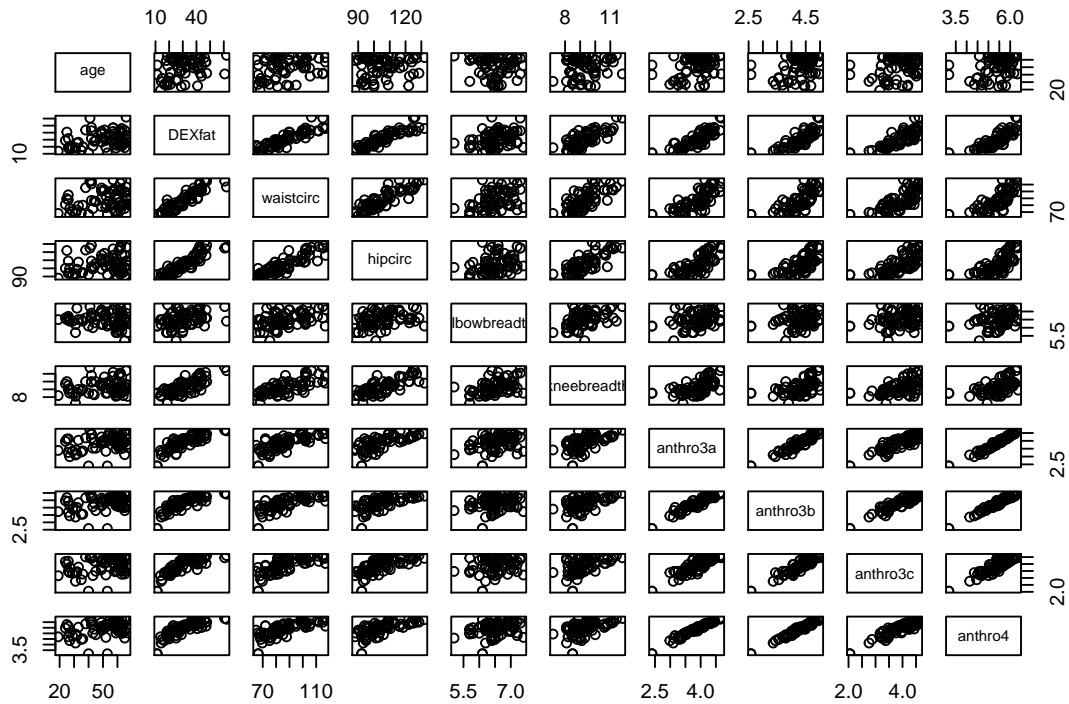
Packages: HSAUR3, mgcv, GGally, mboost, rpart, wordcloud, ggplot2, TH.data, tidyverse, gamair

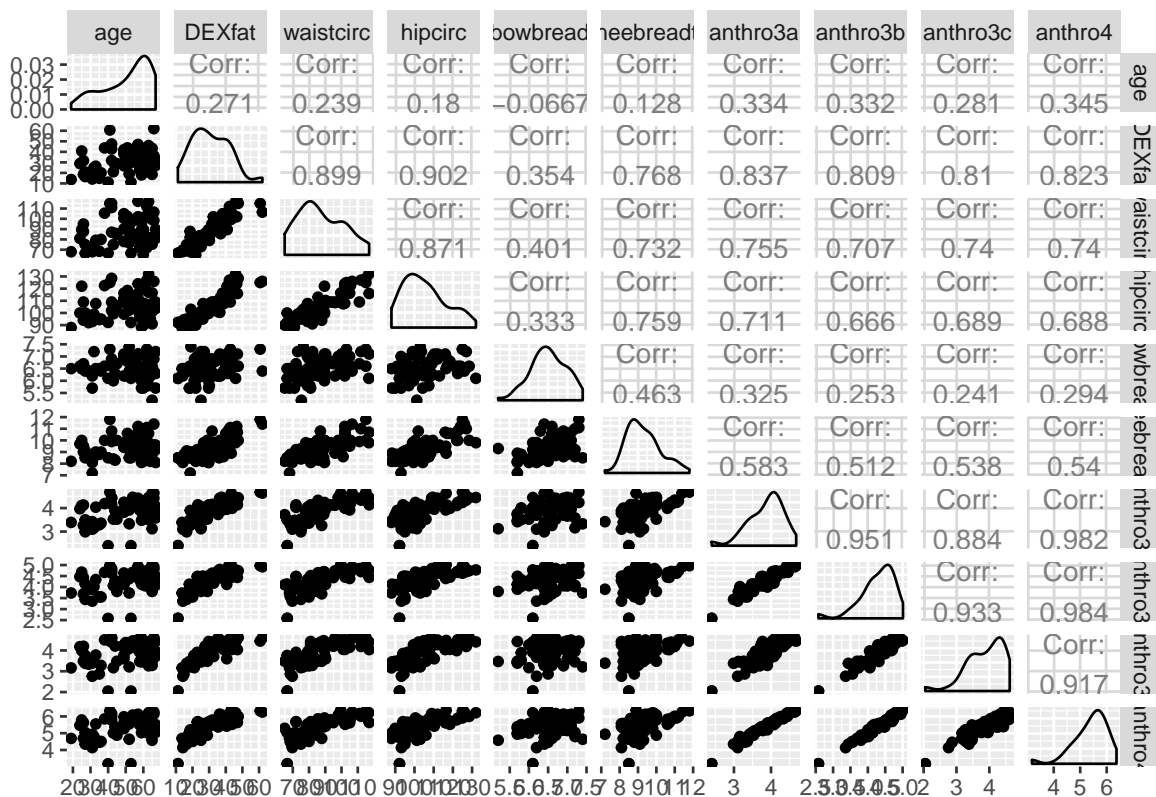
Collaborators: Ajay Gupta, James Young

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 because of the negative impact that highly correlated variables can have. Anthro4 is highly correlated with anthro3a and anthro3b and anthro3a is also highly correlated with anthro3b.

```
## Features selected after removing variables correlated greater than 0.94:
```

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

```
##
```

```
## Features selected after removing variables correlated greater than 0.80:
```

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

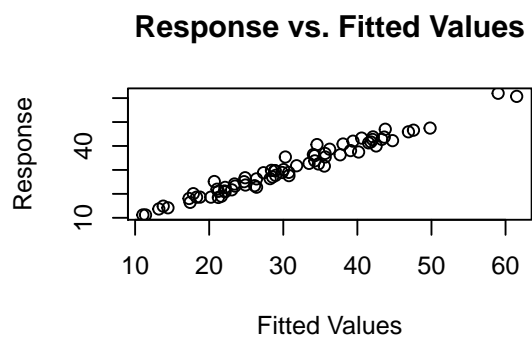
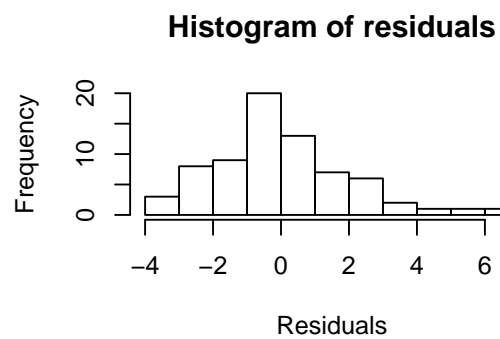
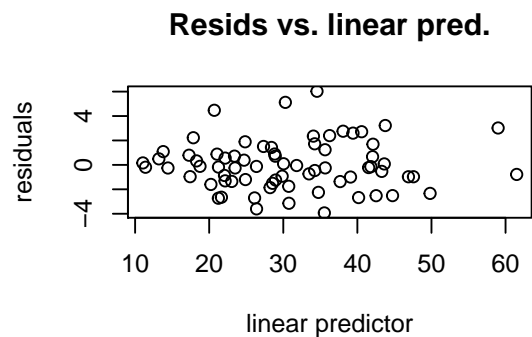
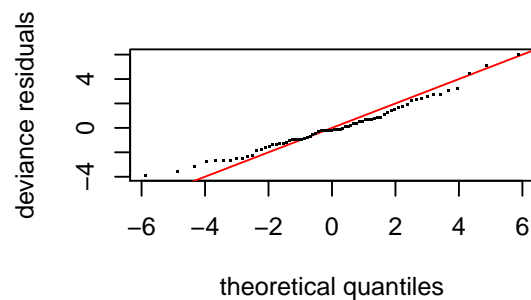
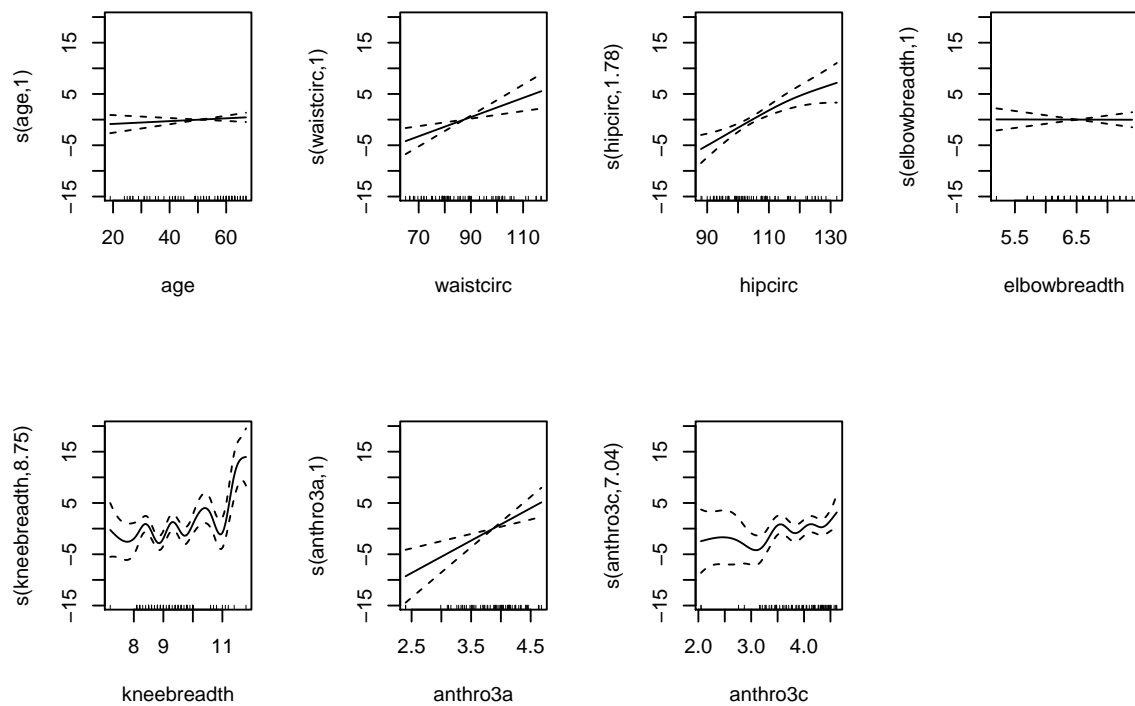
Discussion - Since we want to have both waistcirc, anthro3c, and hipcirc in part b, the threshold for correlation elimination was set to 0.94 to remove those that are highly correlated while maintaining these 3 features (anthro3c has a correlation of 0.933 with anthro3b, so I set the threshold slightly 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*. If the threshold was set lower to remove more highly correlated values (correlations greater than 0.8), I think *age*, *waistcirc*, *elbowbreadth*, *kneebreadth*, and *anthro3a* would be important to use without having built any model yet.

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-R², 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
```



```
##
## 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
##
## Model GVC:
## [1] 8.435412
##
## Model AIC:
## [1] 345.708
##
## Model R^2:
## [1] 0.9528156
##
## Estimated Degrees of Freedom:
## [1] 21.57091

```

Discussion - GCV = 8.44, AIC = 345.71, $\text{adj-R}^2 = 0.95$, 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 21.57 degrees of freedom. The GCV and AIC most likely could be improved through feature selection than smoothing those features in the generalized additive model, but without another model to compare to it cannot be determined if this is a good model or poor model. Using *gamcheck()* to look at the residual plots, the assumption of normality doesn't appear to be violated. The histogram shows that a couple residuals are greater than 4, but as a whole do not appear to violate the assumption.

- 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 +

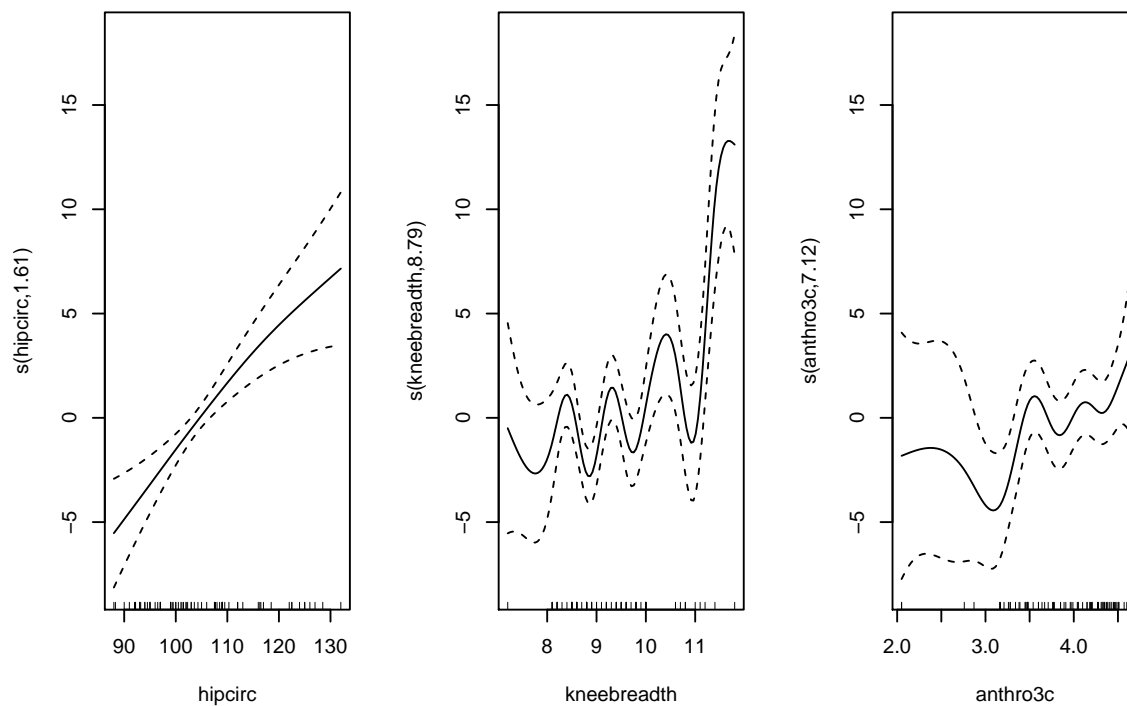
```

```

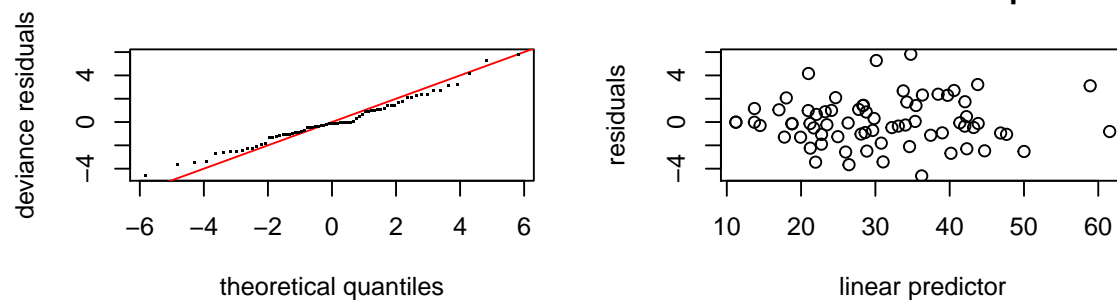
s(anthro3c), data = bodyfat)

## 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

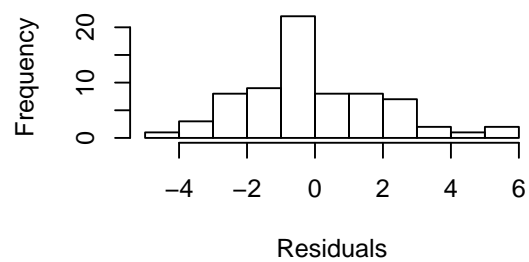
```



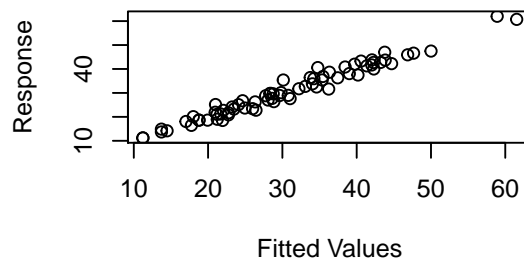
Resids vs. linear pred.



Histogram of residuals



Response vs. Fitted Values



```
##
## 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  0.47
## s(kneebreadth) 9.00 8.79    1.06  0.68
## s(anthro3c)   9.00 7.12    0.91  0.23
##
## Model GVC:
## [1] 7.946447
##
## Model AIC:
## [1] 343.2562
##
## Model R^2:
## [1] 0.9536683
##
## Estimated Degrees of Freedom:
## [1] 19.52001

```

Discussion - GCV = 7.95, AIC = 343.26, adj-R² = 0.95. All variable terms were significant at $\alpha = 0.05$ and the intercept was significant at $\alpha = 0.1$ 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-R² 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-R², residual plots, etc. to compare models).

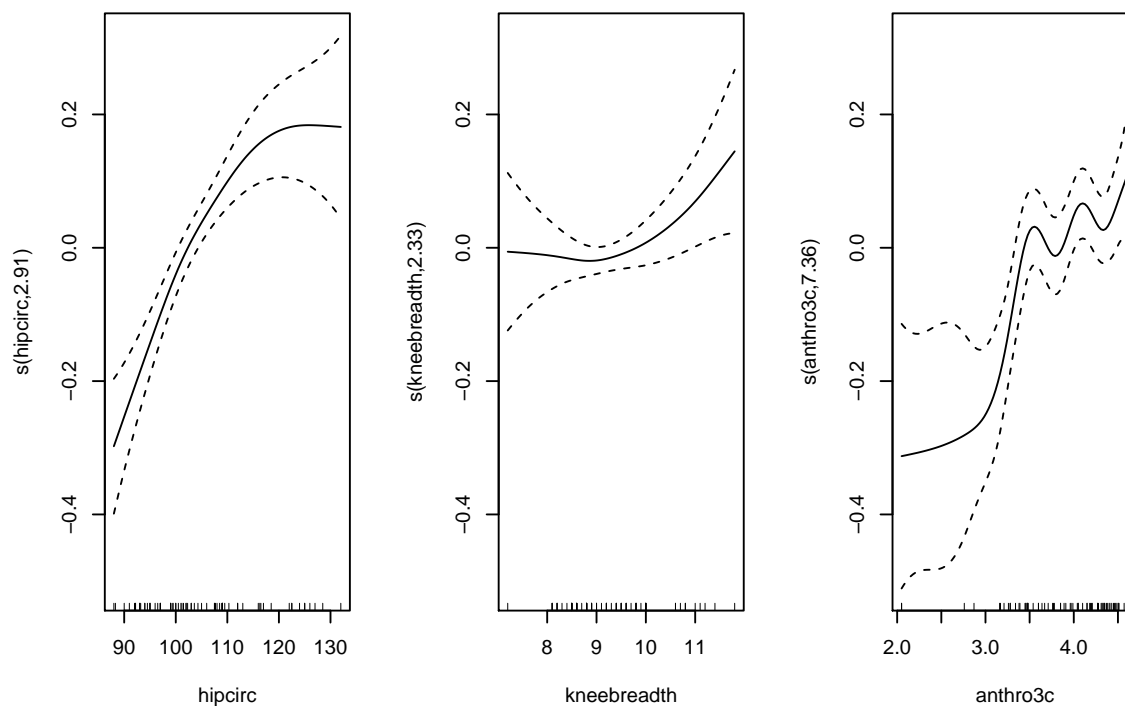
```

##
## 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

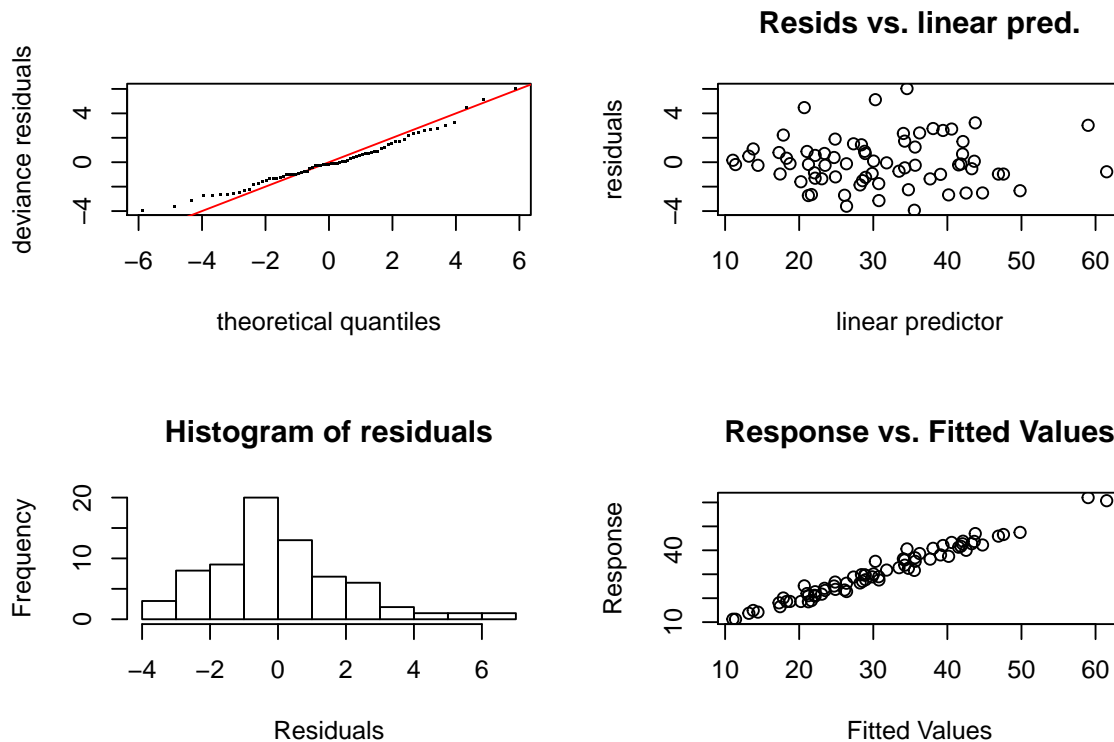
```



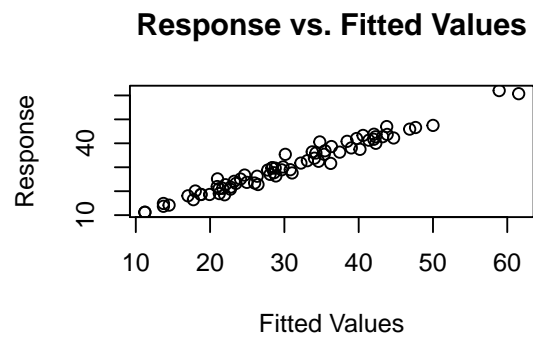
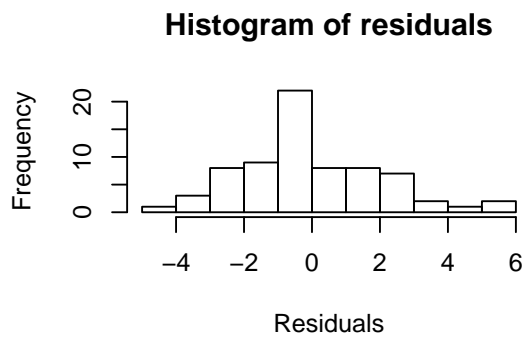
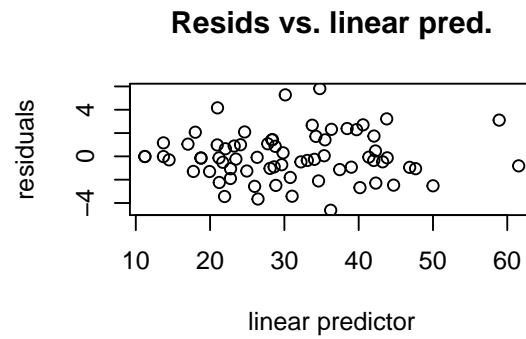
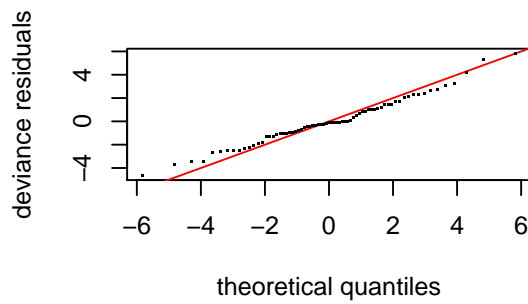
```
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(hipcirc)  2.909  3.616 11.828  8.8e-07 ***
## s(kneebreadth) 2.325  2.962  2.027 0.128320
## s(anthro3c)  7.358  8.263  4.678 0.000144 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.952   Deviance explained = 96.2%
## GCV = 0.0088137   Scale est. = 0.006878   n = 71
```



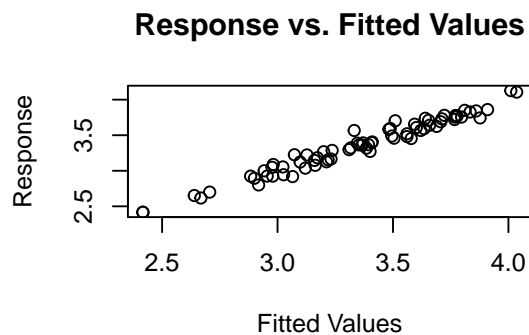
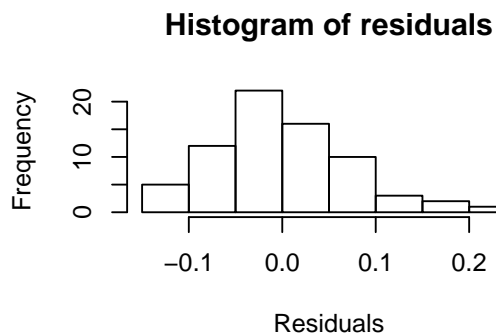
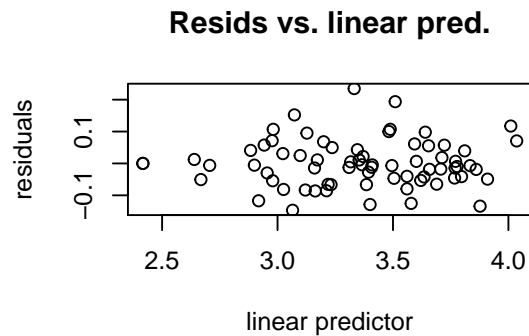
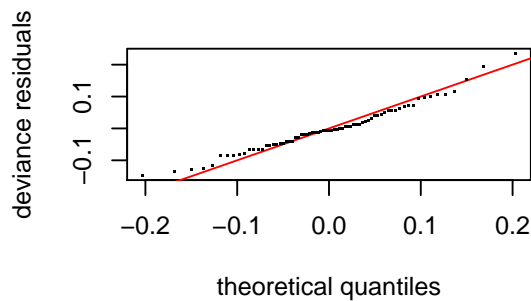
```
##
## Part B Model:
```



```
##
## 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
##
## Model GVC:
## [1] 0.008813659
##
## Model AIC:
## [1] -136.47
##
## Model R^2:
## [1] 0.9522733
```

```
##
## Estimated Degrees of Freedom:
## [1] 14.59274
```

Discussion - The log transformed response variable decreases the generalized cross-validation score and the AIC of the model which indicates that log transforming the response variable creates a better model than not log transforming it. 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 $\alpha=0.05$ now too. The degrees of freedom was also decreased from smoothing using the log transformation of the same model as Part C from an estimated 19.5 to 14.6 DoF. The adj-R^2 was slightly less than non-log transformed DEXfat, but only 0.001 less.

- 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
```

Discussion - 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.

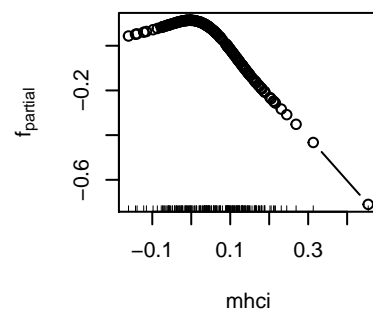
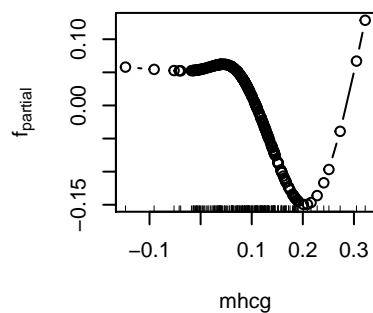
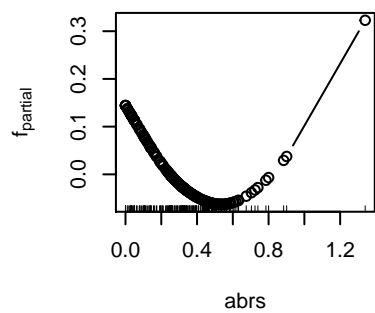
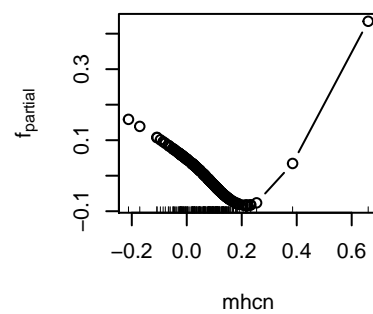
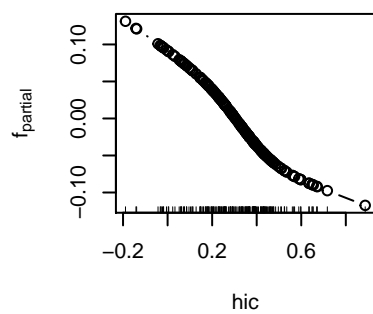
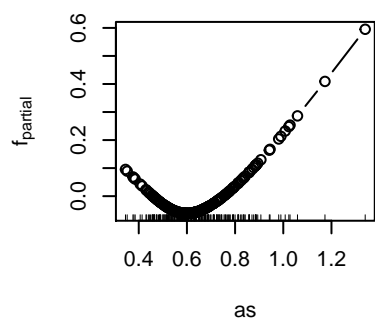
- Fit a logistic additive model to the glaucoma data. (Here use family = "binomial"). Which covariates should enter the model and how is their influence on the probability of suffering from glaucoma? (Hint: since there are many covariates, use `gamboost()` to fit the GAM model.)

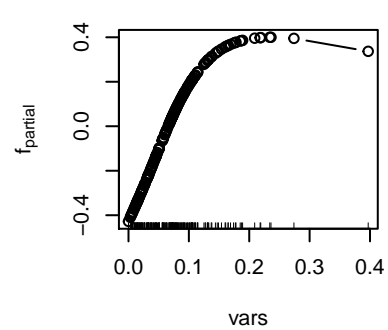
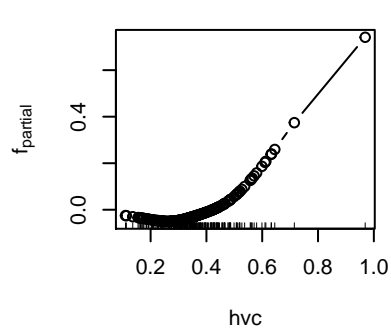
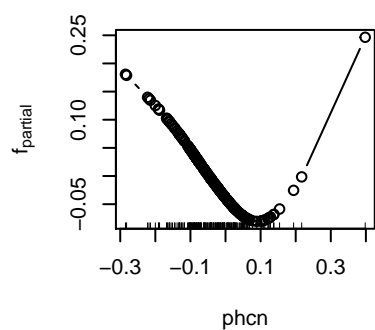
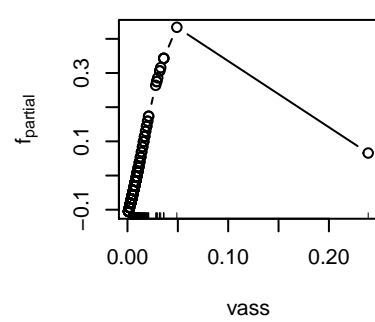
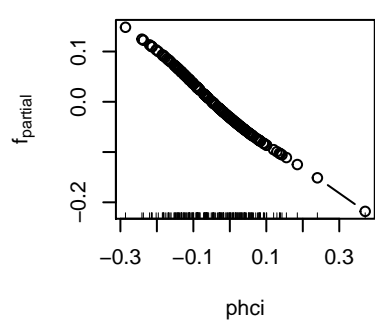
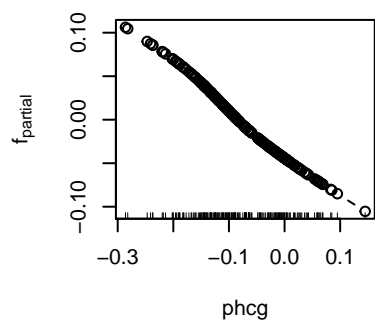
```
##
```

```

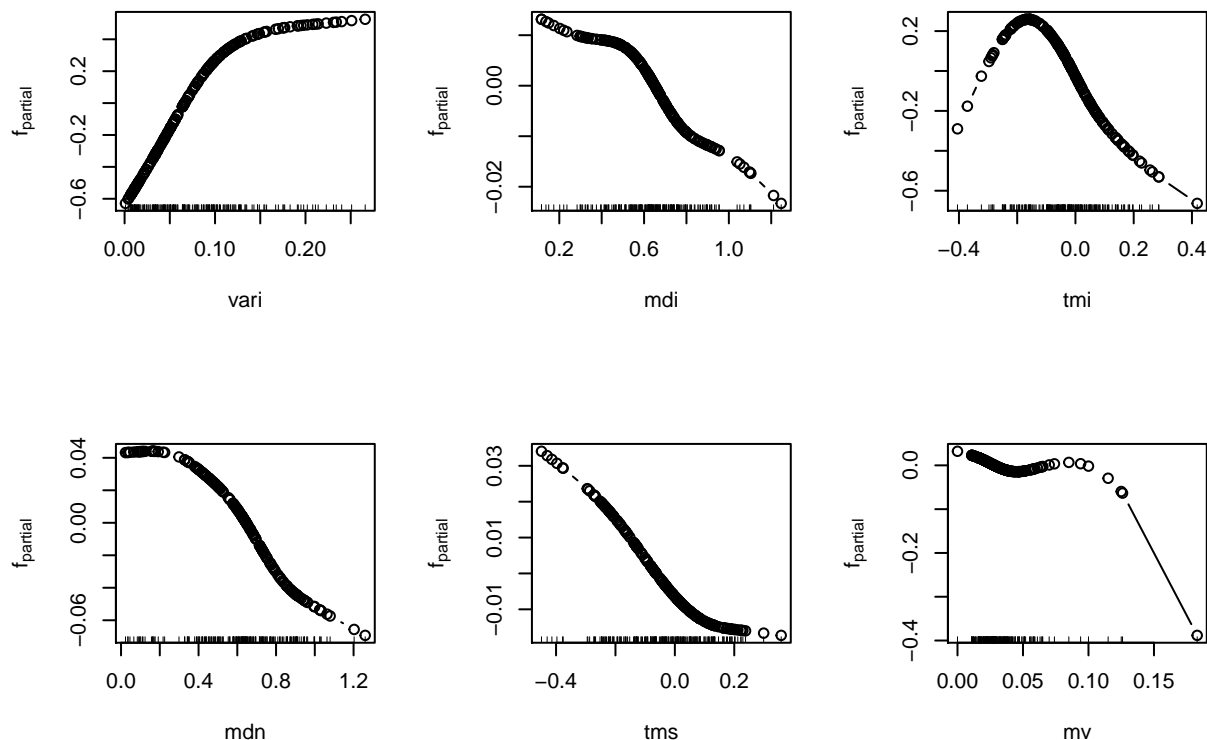
Model-based Boosting
##
Call:
gamboost(formula = Class ~ ., data = glau, family = Binomial())
##
Negative Binomial Likelihood (logit link)
##
Loss function: {
f <- pmin(abs(f), 36) * sign(f)
p <- exp(f)/(exp(f) + exp(-f))
y <- (y + 1)/2
-y * log(p) - (1 - y) * log(1 - p)
}
##
##
Number of boosting iterations: 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
##
Error rate of gamboost fitted model:
[1] 0.1020408

```









```
##
Family: binomial
Link function: logit
##
Formula:
obs ~ s(as, k = 3) + hic + s(mhcn, k = 3) + s(abrs, k = 3) +
s(mhcg, k = 3) + s(mhci, k = 3) + phcg + phci + s(vass, k = 3) +
s(phcn, k = 3) + s(hvc, k = 3) + s(vars, k = 3) + s(vari,
k = 3) + mdi + s(tmi, k = 3) + s(mdn, k = 3) + s(tms, k = 3) +
s(mv, k = 3)
##
Parametric coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 40.084 14.563 2.752 0.00592 **
hic 31.699 13.415 2.363 0.01813 *
phcg 39.616 22.473 1.763 0.07794 .
phci -26.497 13.544 -1.956 0.05042 .
mdi -12.350 5.504 -2.244 0.02484 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Approximate significance of smooth terms:
edf Ref.df Chi.sq p-value
s(as) 1.815 1.965 6.040 0.03329 *
s(mhcn) 1.000 1.000 0.404 0.52500
s(abrs) 1.000 1.000 4.763 0.02908 *
```

```

s(mhcg) 1.000 1.000 3.031 0.08170 .
s(mhci) 2.000 2.000 13.591 0.00112 **
s(vass) 2.000 2.000 12.748 0.00173 **
s(phcn) 1.816 1.963 2.879 0.20468
s(hvc) 1.904 1.989 9.753 0.00559 **
s(vars) 1.000 1.000 0.264 0.60708
s(vari) 1.879 1.984 4.422 0.11529
s(tmi) 2.000 2.000 11.734 0.00284 **
s(mdn) 1.000 1.000 2.111 0.14624
s(tms) 1.001 1.001 0.485 0.48636
s(mv) 1.383 1.619 4.129 0.05556 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
R-sq.(adj) = 0.816 Deviance explained = 80.3%
UBRE = -0.46313 Scale est. = 1 n = 196

##
MSE of rebuilt Model using variables given by gamboost:
[1] 0.03571429

##
Error rate of rebuilt model:
[1] 0.03571429

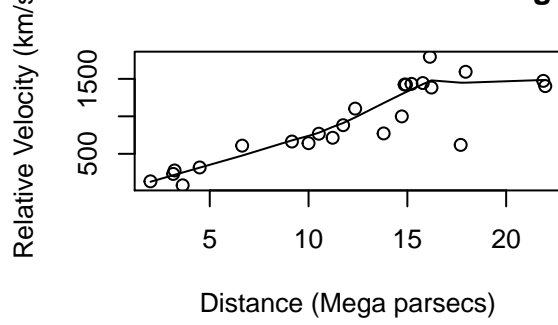
Class_cal
glaucoma normal
glaucoma 95 3
normal 4 94

```

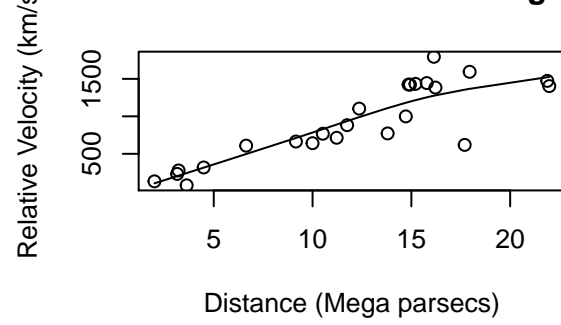
Discussion - The variable selected using *gamboost()* are *tmi*, *mhcg*, *vars*, *mhci*, *hvc*, *vass*, *as*, *vari*, *mv*, *abrs*, *mhcn*, *phcn*, *mdn*, *phci*, *hic*, *phcg*, *mdi*, and *tms*. Using these variables, the classification error of the model (error and MSE = 0.102) is lower than that of 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). Creating another model to allow for a nice summary of the variables, we can see that *mhcn*, *phcn*, *vars*, *vari*, *mdn* and *tms* aren't significant.

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

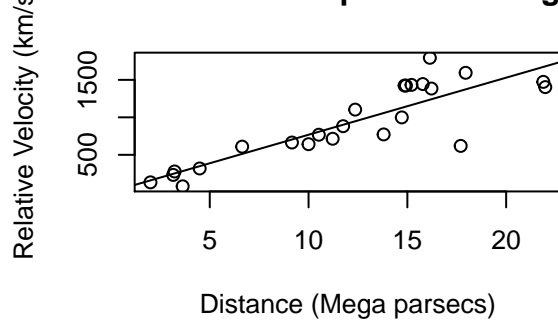
**Scatter Plot with Lowess Smoothing Function**



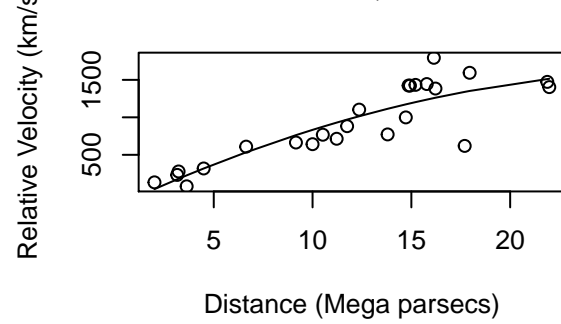
**Scatter Plot with Cubic Smoothing function**

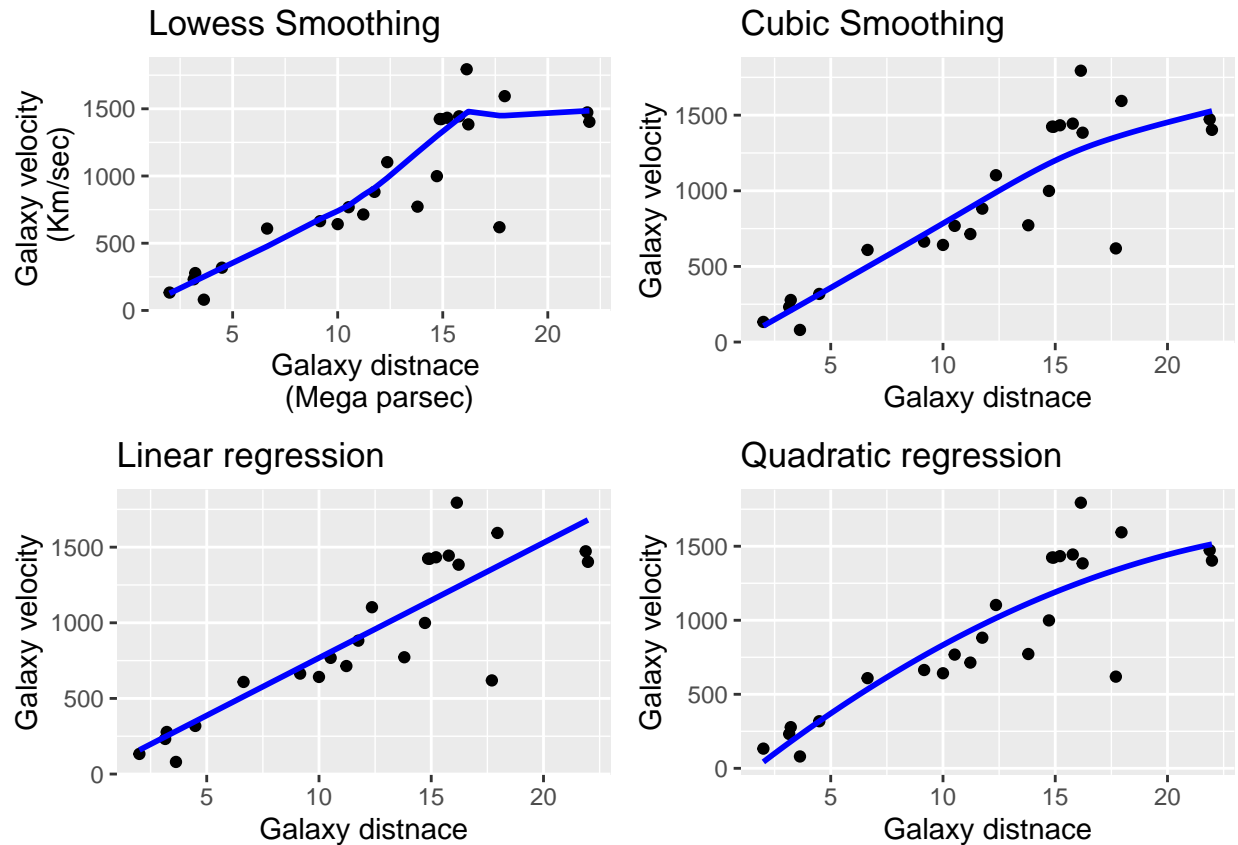


**Scatter Plot with Simple Linear Regression**



**Scatter Plot with Quadratic Model**





Discussion - I used 4 different scatter plot models with the hubble data using the example from the handbook. The linear model is as expected and previously seen in homework. The Lowess smoothing shows a couple seemingly sharp changes in direction while the quadratic and cubic smoothing models are less abrupt while still appearing to fit the data decently.

Resources Used:

- [stat.ethz.ch](http://stat.ethz.ch)
- [rdocumentation.org](http://rdocumentation.org)
- [stackexchange.com](http://stackexchange.com)
- Homework 5