Homework 6

STAT 601

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

Packages: TH.data, GGally, ggplot2, mgcv, dplyr, tidyr, mboost, gamair

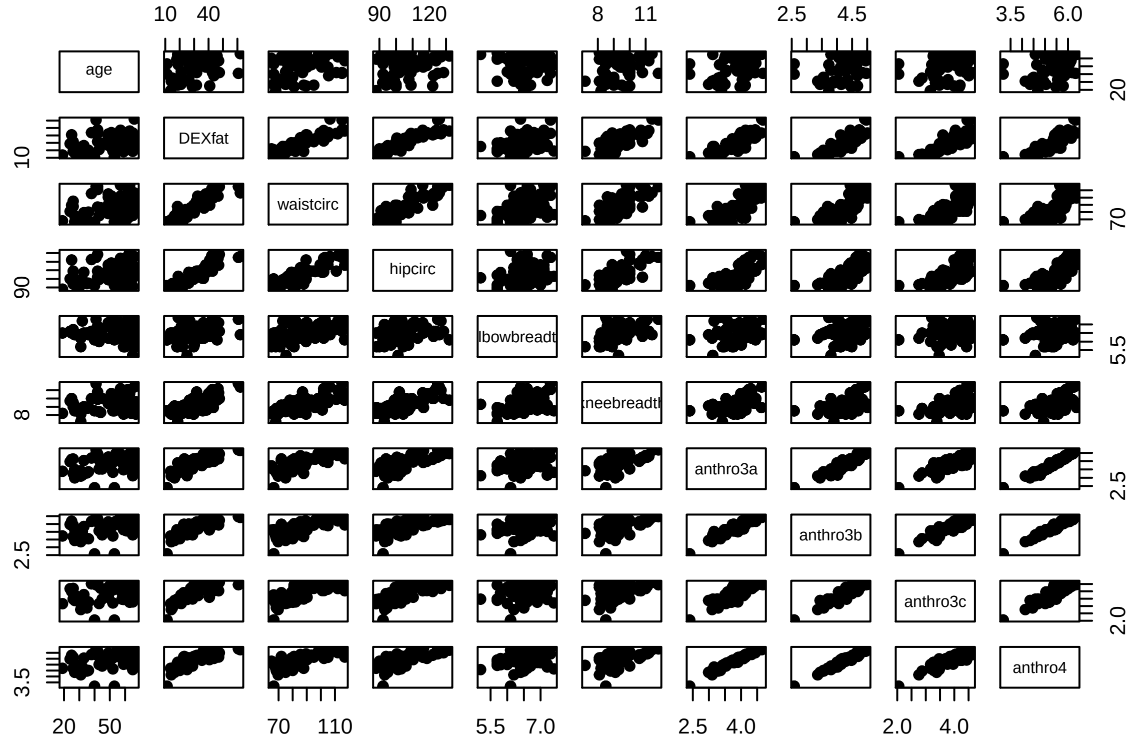
Collaborators: Alex Soupir, Ajay Gupta

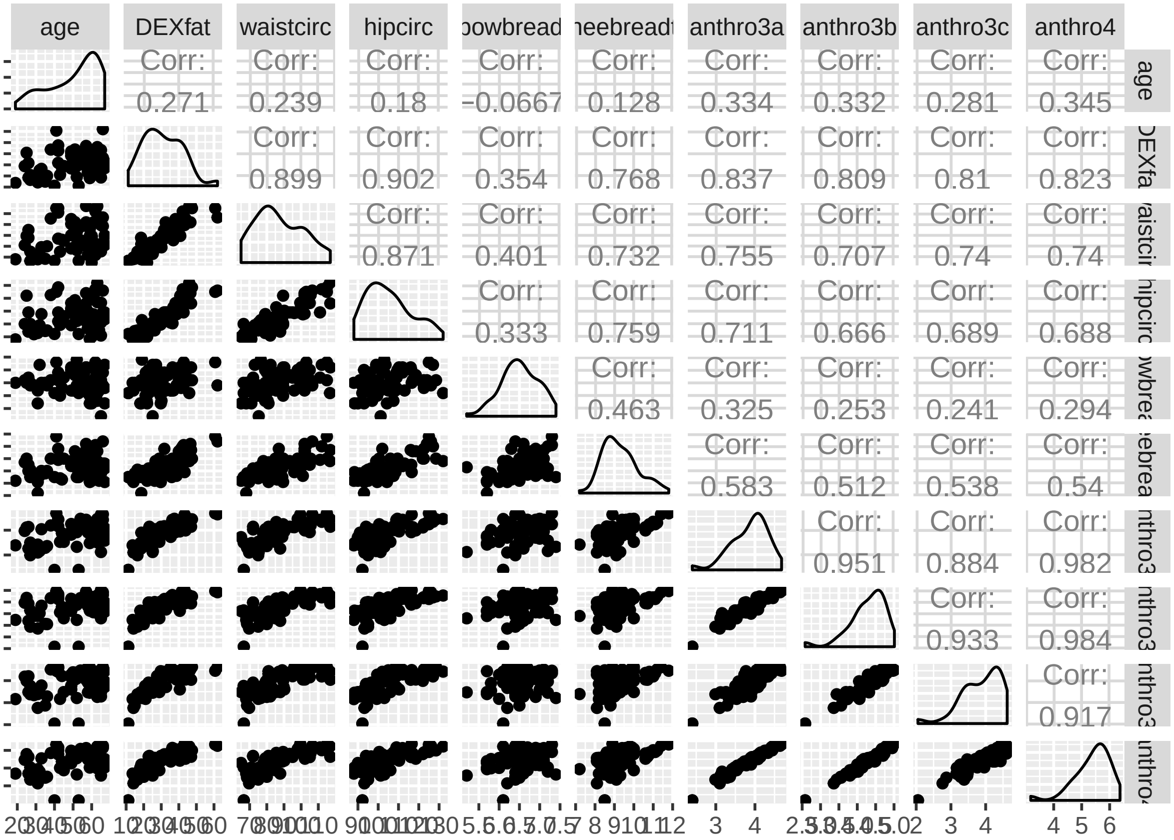
Resources: Stack Overflow, https://www.reddit.com/r/AskStatistics/comments/5ydt2c/if_my_aic_and_bic_are_negative_does_that_mean/ (interpreting negative AIC)

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).

Below I have created a visualization of variable correlation using base r and ggpairs. DEXfat is the variable we will be predicting later so variables showing correlation with DEXfat suggest they may be good input variables. We can see below that variables with the prefix "anthro" have strong correlations with each other. Another variable pair that shows correlation is waist circularity and hip circularity.





Since we are told below what variables to use in the model, we do not need to remove variables right now based on correlation.

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

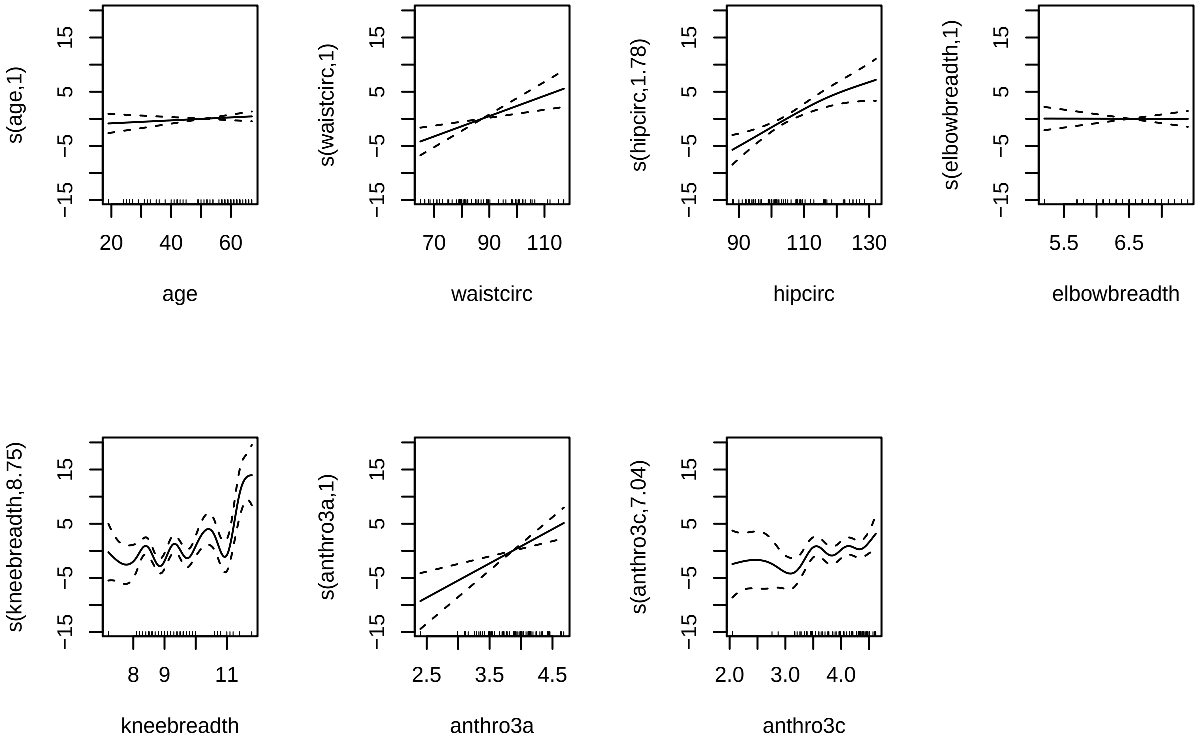
```
bodyfat_gam <- gam(DEXfat~ s(age) + s(waistcirc) + s(hipcirc) +
          s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) +
          s(anthro3c), data = bodyfat)
```

- 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?

In my opinion, not all covariates are informative based on the given plots, more specifically, age is not informative visually and the sumamry says elbow breadth is not informative (p-value ~0.9). Some of the variables could be included as linear effects based on these plots, specifically waist circularity, elbow breadth, and anthro3a.

```
##
## 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
```



- Report GCV, AIC, adj-R^2, and total model degrees of freedom.

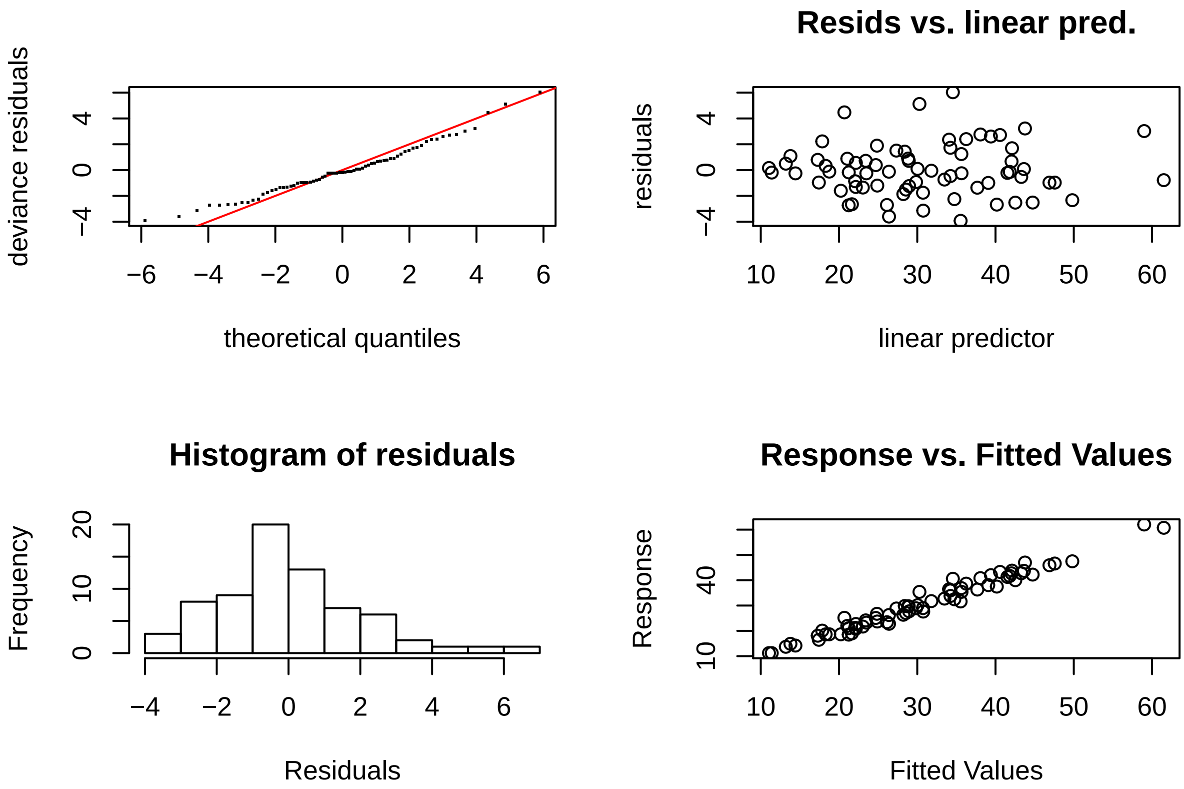
Below we see the given AIC, GCV, and DF which will be more meaningfull when compared

with other models, but for right now the adjusted R-squared of ~0.95 is a hint that this is a pretty good model, especially as a baseline.

```
##          gam.stat
## GCV         8.435
## AIC       345.708
## Adj.Rsq     0.953
## DF         21.571
```

- Use **gam.check()** function to look at the diagnostic plot. Does it appear that the normality assumption is violated?

Below we see diagnostic plots that demonstrate a fairly normal distribition of residuals, albeit with a slightly long tail to the right on the histogram. It's not perfect, but visually it looks close to normal.



```
##
## 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
```

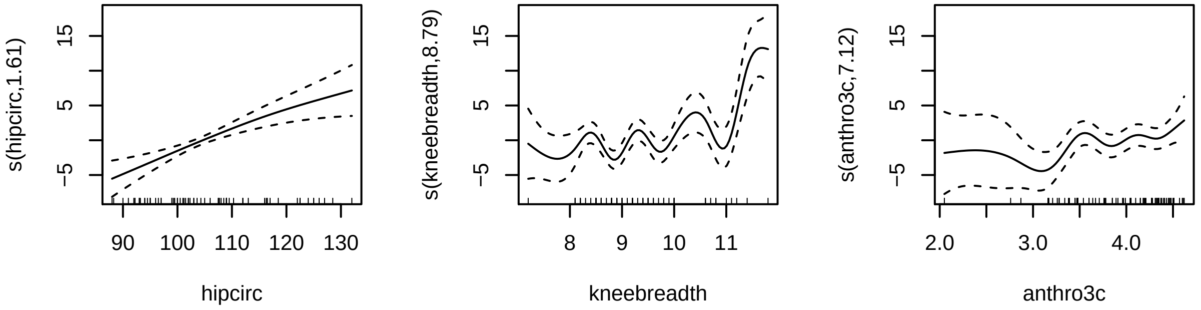
- Write a discussion on all of the above points.

We started by constructing a generalised additive model (GAM) with prescribed variables to predict DEXfat. We then visualized the relations between the input variables and output variable as determined by the GAM. Some variables had a linear relation, others had non-linear relations, and age appeared to have no relation visually. We also looked at the summary and saw that age, elbow breadth, and anthro3c were not significant at alpha = 0.05 in this model. I made a report table of the AIC, GCV, Adj. R-squared, and degrees freedom. The adjusted R-squared of ~0.95 suggests this is a good model. Finally I looked at diagnostic plots and viewed the residuals which seem close to a normal distribution, but not perfect as stated above. verall, this model seems like a good start to me and now we can tweak to look for improvement.

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

```
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
```



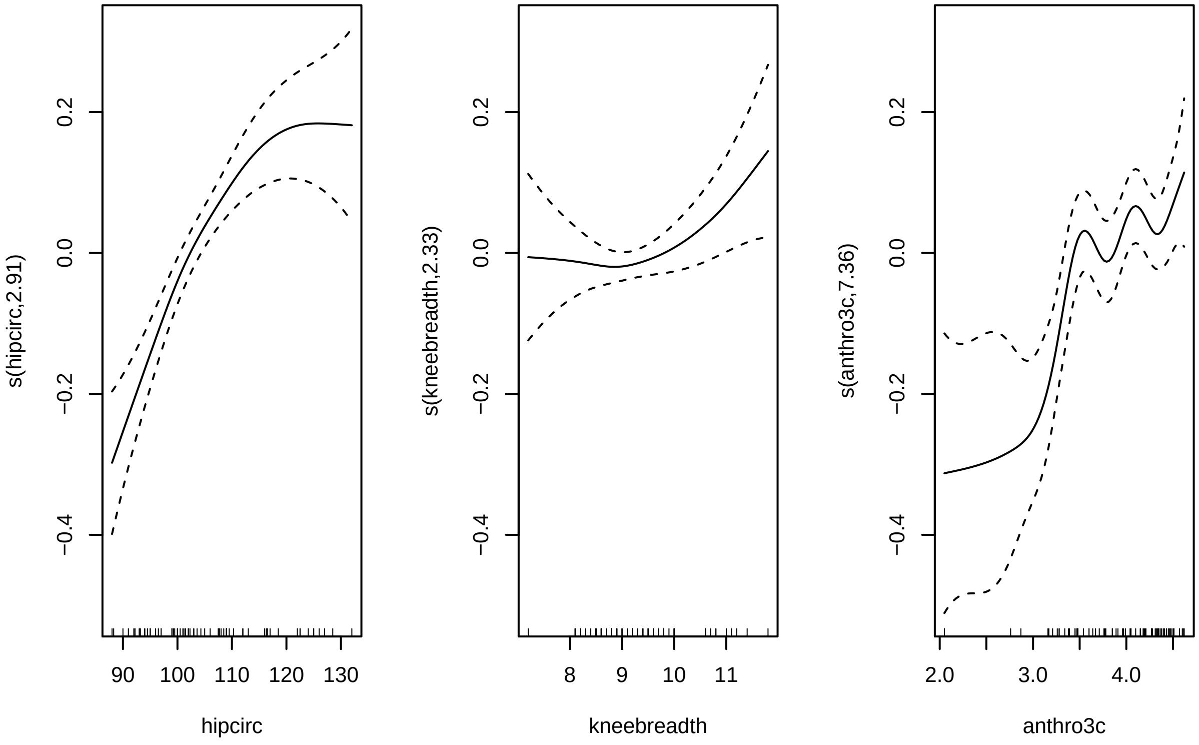
Below we see a slightly lower GCV and AIC with an adjusted R-square that is 0.01 higher suggesting this model is better than the first model.

```
##         gam2.stat
## GCV         7.946
## AIC       343.256
## Adj.Rsq     0.954
## DF         17.520
```

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).

I fit the model and displayed plots of the smoothed functions below as well as this models goodness of fit indicators. Further below I compare all three 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
```



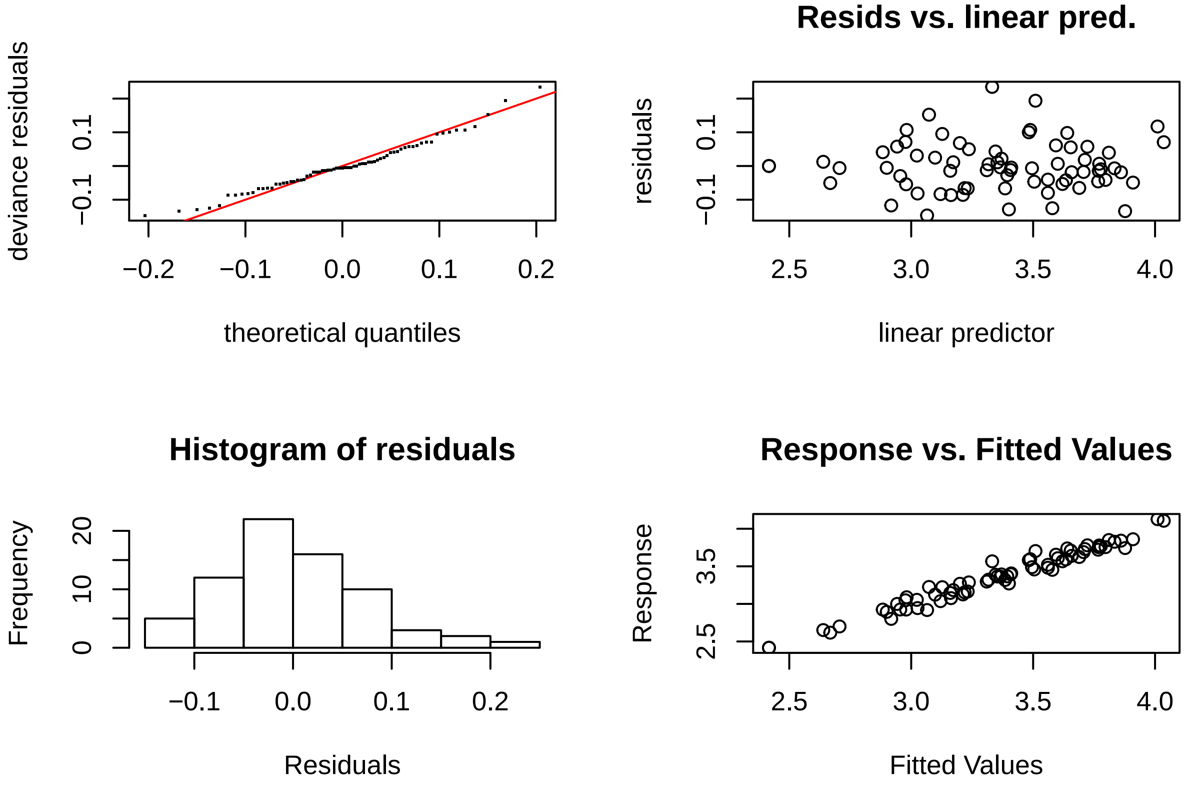
```
##          gamlog.stat
## GCV            0.009
## AIC         -136.470
## Adj.Rsq        0.952
## DF            12.593
```

Compare The Three Models

Below we can see that all 3 models have adjusted R-squared values of ~0.95 with a range of 0.002 from highest to lowest. The lowest R-squared model also has the lowest AIC and degrees of freedom, meaning a simpler model is delivering almost the same predictive power.

```
##         gam.stat gam2.stat gamlog.stat
## GCV        8.435     7.946       0.009
## AIC      345.708   343.256    -136.470
## Adj.Rsq    0.953     0.954       0.952
## DF        21.571    17.520      12.593
```

checking the residuals, we again see that that the log transformed GAM model does not have a perfect normal distribution of residuals, in fact it looks pretty similar to the first model. However, with its similar performance and better AIC due to less parameters, I still would say this log transformed model is the more appropriate choice.



```
##
## 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    0.09 .
## s(kneebreadth) 9.00 2.33    0.83    0.07 .
## s(anthro3c)    9.00 7.36    0.99    0.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)]
```

Below we can see the only variable that was removed using this method was age, which was the only variable I said seemed visually uninformative at the beginning of this work.

```
##
##   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
```

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.)

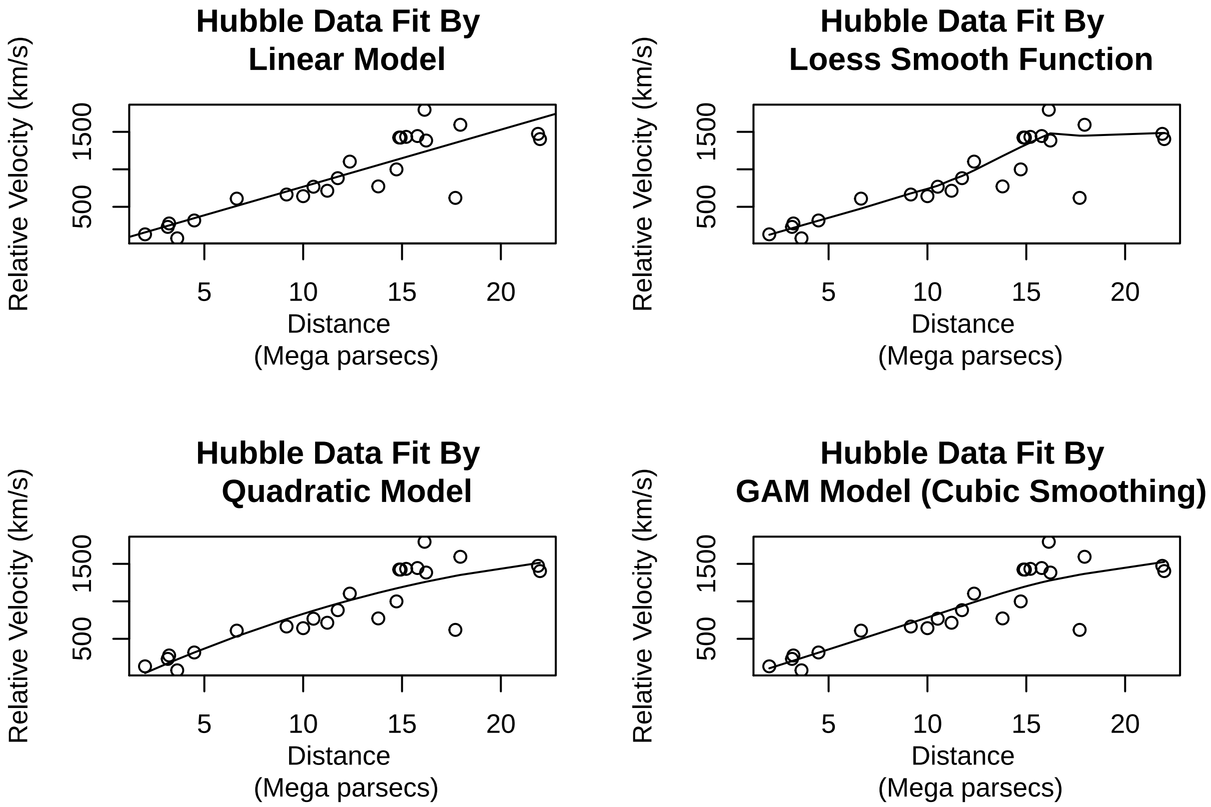
```
##
##   Model-based Boosting
##
## Call:
## gamboost(formula = Class ~ ., data = GlaucomaM, family = Binomial())
##
##
##   Negative Binomial Likelihood (logit link)
##
## Loss function: {
##      f <- pmin(abs(f), 36) * sign(f)
##      p <- exp(f)/(exp(f) + exp(-f))
##      y <- (y + 1)/2
##      -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
```

Using gamboost() to select variables resulted in the selection of tmi, mhcg, vars, mhci, hvc, vass, as, vari, mv, abrs, mhcn, phcn, mdn, phci, hic, phcg, mdi, and tms. This was much more convenient in my opinion than last weeks method of getting rid of covariates.

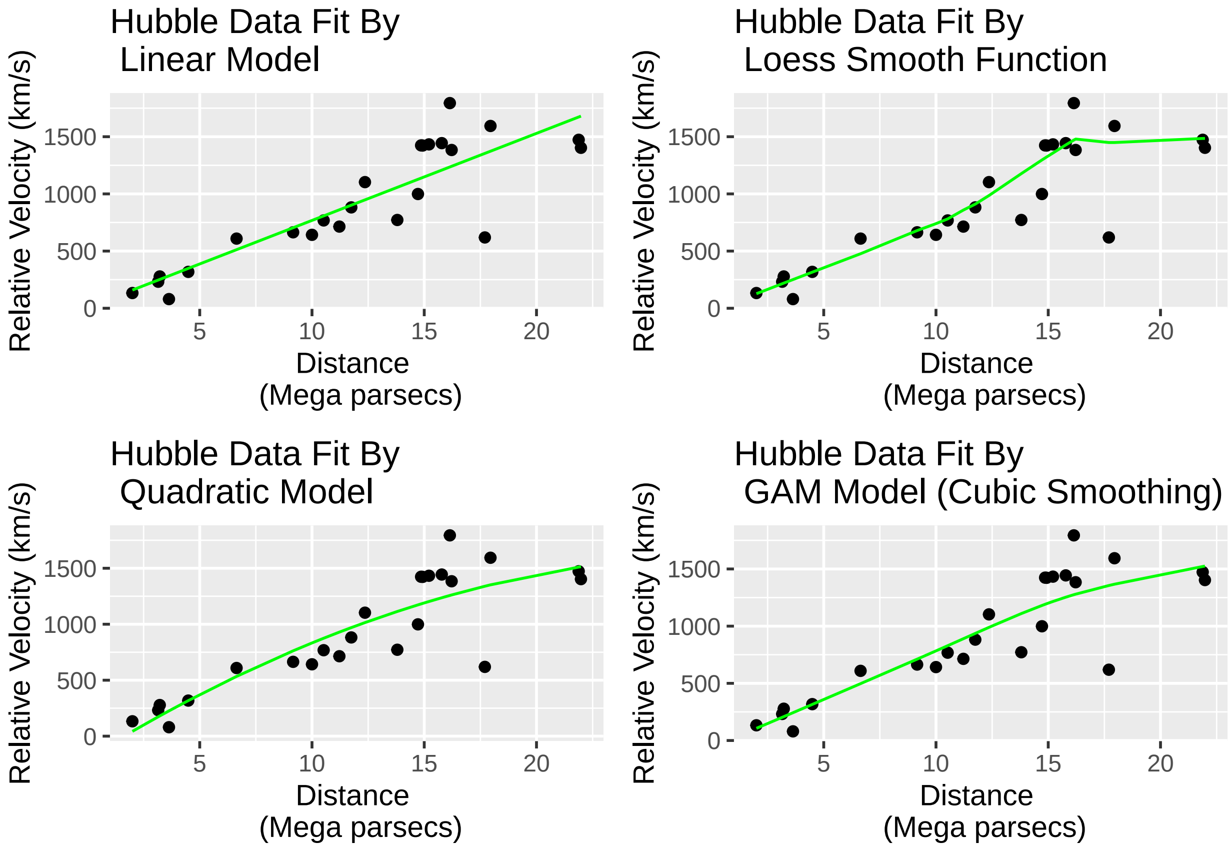
3. Investigate the use of different types of scatterplot smoothers on the Hubble data from Chapter 6. (Hint: follow the example on men1500m data scattersmoothers page 199 of Handbook).

Below we see the scatterplots made with base R and ggplot2 which demonstrate how different smoothing functions fit the data. The quadratic and gam appear visually similar with the linear and lowess appearing distinctly different. This is just

Base R scatterplots



GGPlot scatterplots



Done.