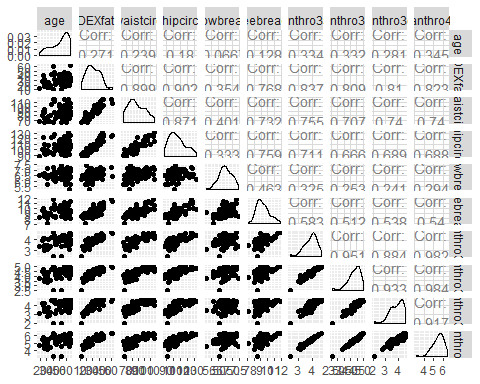
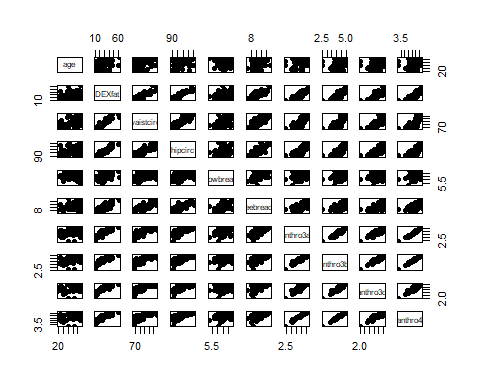
Homework 6

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

1. Consider the body fat data introduced in Chapter 9 ( data from package).
   1. Explore the data graphically. What variables do you think need to be included for predicting bodyfat? (Hint: Are there correlated predictors).

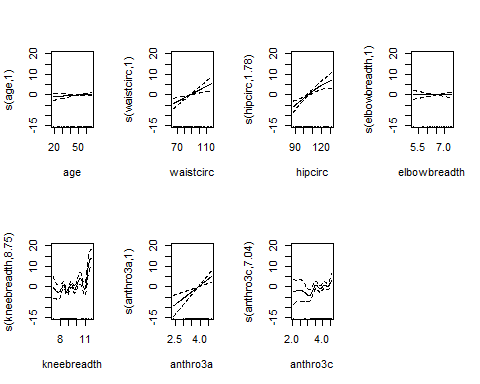


b) Fit a generalised additive model assuming normal errors using the following code.   
  
 (\begin{verbatim}  
 bodyfat\_gam <- gam(DEXfat~ s(age) + s(waistcirc) + s(hipcirc) +   
 s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) +  
 s(anthro3c), data = bodyfat)  
 \end{verbatim})

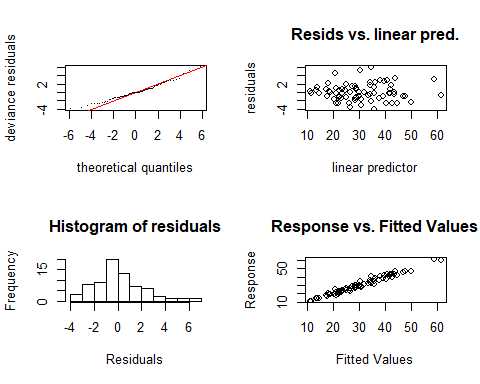
bodyfat\_gam <- gam(DEXfat~ s(age) + s(waistcirc) + s(hipcirc) + s(elbowbreadth) + s(kneebreadth)+ s(anthro3a) + s(anthro3c), data = bodyfat)

- Assess the \textbf{summary()} and \textbf{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?

##   
## 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



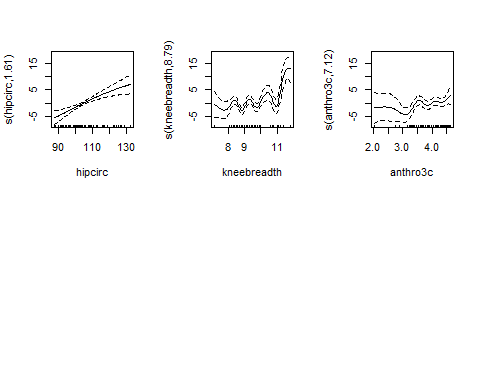
- Report GCV, AIC, adj-R$^2$, and total model degrees of freedom.  
  
  
  
 - Use \textbf{gam.check()} function to look at the diagnostic plot. Does it appear that the normality assumption is violated?



##   
## 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

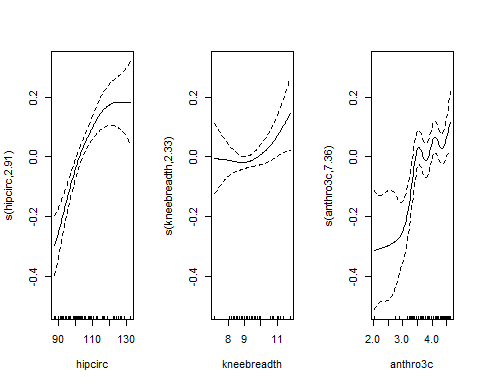
- Write a discussion on all of the above points.  
   
  
  
  
  
  
  
c) Now remove insignificant variables and remove smoothing for some variables. Report the summary, plot, GCV, AIC, adj-R$^2$.  
   
 \begin{verbatim}  
 bodyfat\_gam2 <- gam(DEXfat~ waistcirc + s(hipcirc) +   
 s(kneebreadth)+ anthro3a +  
 s(anthro3c), data = bodyfat)  
 \end{verbatim}

##   
## 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



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



e) Fit a generalised additive model that underwent AIC-based variable selection (fitted using function \textbf{ gamboost()} function). What variable was removed by using AIC?   
 \begin{verbatim}  
 bodyfat\_boost <- gamboost(DEXfat~., data = bodyfat)  
 bodyfat\_aic <- AIC(bodyfat\_boost)  
 bf\_gam <- bodyfat\_boost[mstop(bodyfat\_aic)]  
 \end{verbatim}

##   
## 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

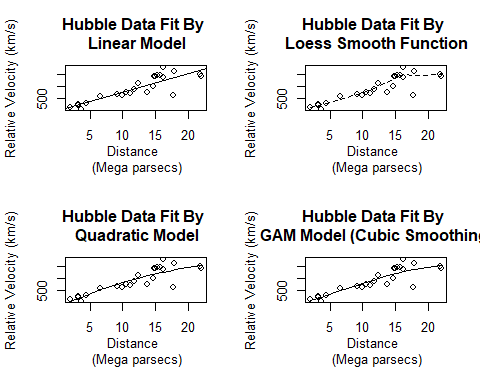
## bbs(waistcirc, df = dfbase) bbs(hipcirc, df = dfbase)   
## "waistcirc" "hipcirc"   
## bbs(elbowbreadth, df = dfbase) bbs(kneebreadth, df = dfbase)   
## "elbowbreadth" "kneebreadth"   
## bbs(anthro3a, df = dfbase) bbs(anthro3b, df = dfbase)   
## "anthro3a" "anthro3b"   
## bbs(anthro3c, df = dfbase)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

1. 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 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.**

1. 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).



Done.