Homework #6

Justin Robinette

October 2, 2018

No collaborators for any problem

Problem #1, Part A: Consider the body fat data introduced in Chapter 9 (bodyfat data from **TH.data** package).

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

Results: First we examined the relationship between predictors using a couple of plots. These are labels *Figure 1.1* and *Figure 1.2*. From this we can see that there are some highly correlated relationships among the predictors. *Figure 1.3* summarizes the correlation values between predictor variables.

To combat this multicollineary, I dropped any variable that had a correlation greater than 0.94 with any other variable. In doing so, we retained 'anthro3c' but dropped 'anthro3b' and 'anthro4' which were both highly correlated with 'anthro3c'.

After removing highly correlated predictors, the variables that should be included for predicting 'DEXfat' are listed in *Figure 1.4*.

Lastly, we look at the correlation between the remaining predictors and the response variable, 'DEXfat', to get a better idea of which predictors will have the biggest affect on the response. This is shown in the graph labelled *Figure 1.5* and summarized in *Figure 1.6*.

Base R plots are included with each ggplot for comparison, per homework guidelines.

Figure 1.1: Bodyfat Correlation Plot 1

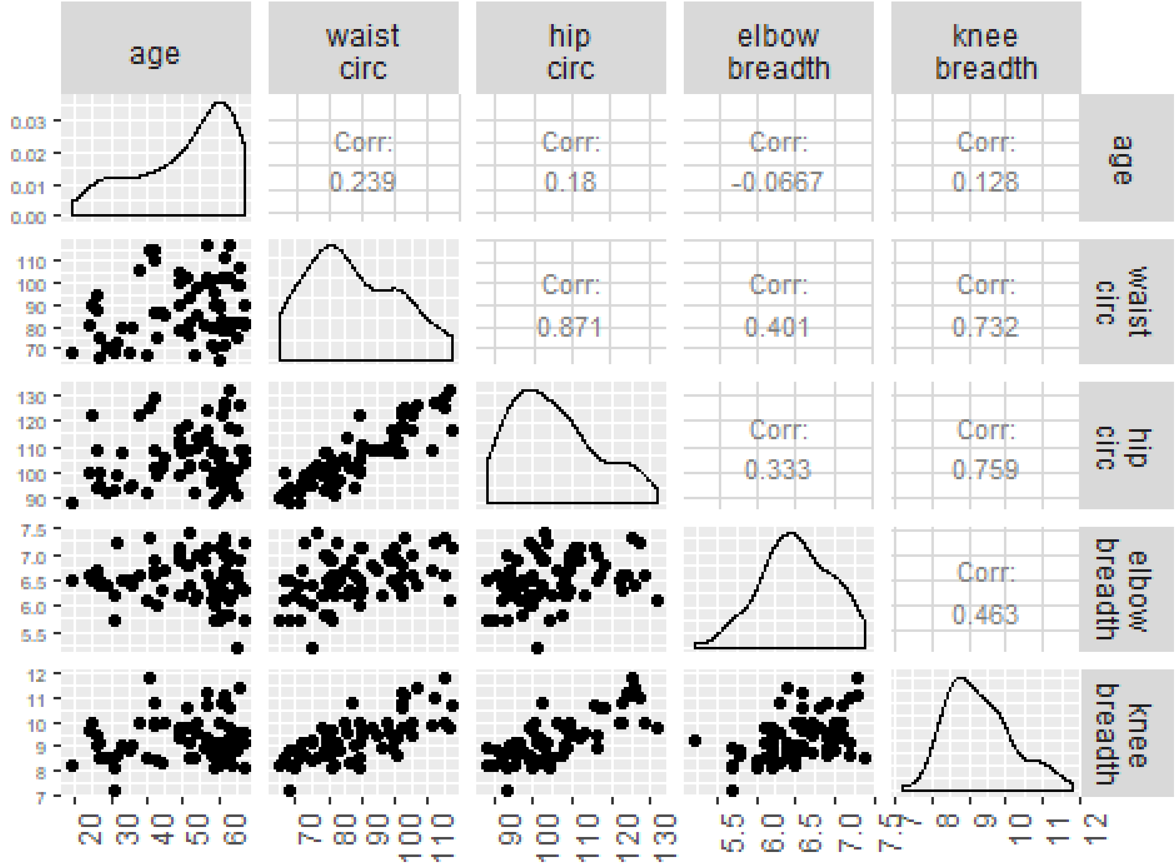
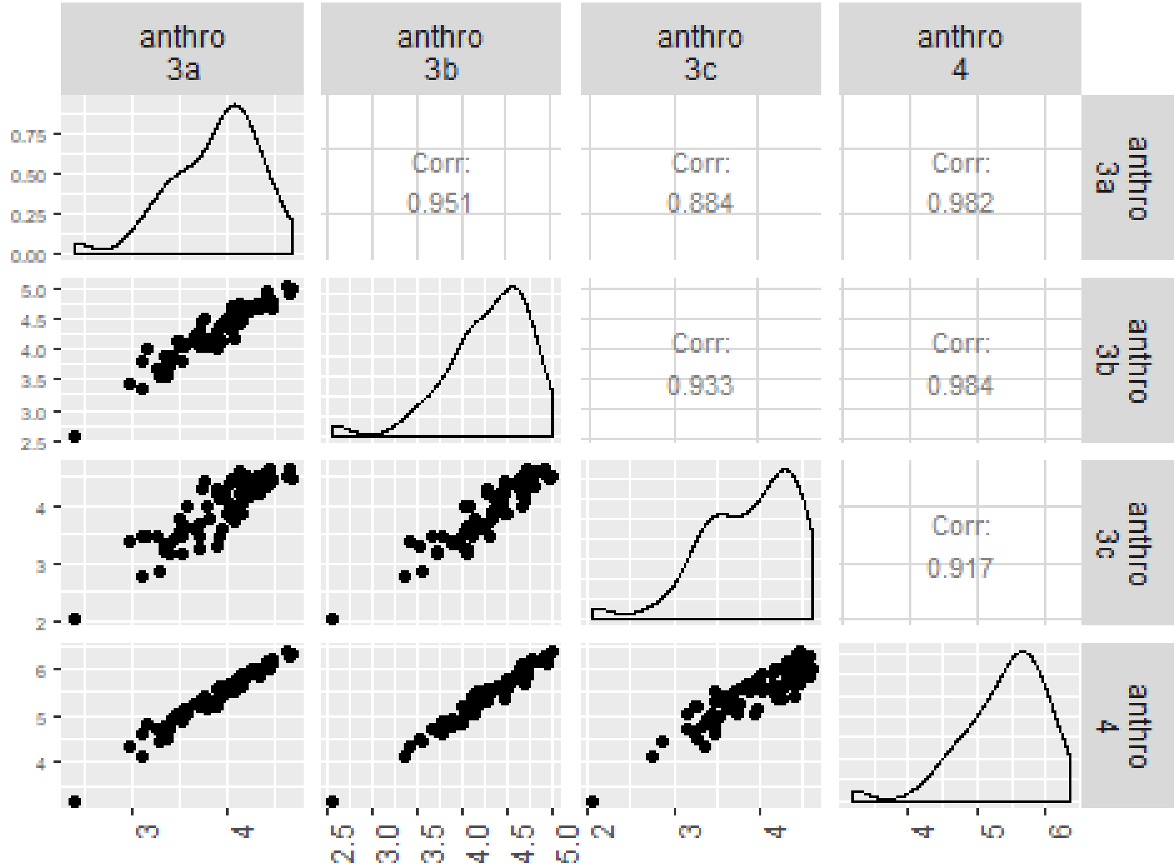
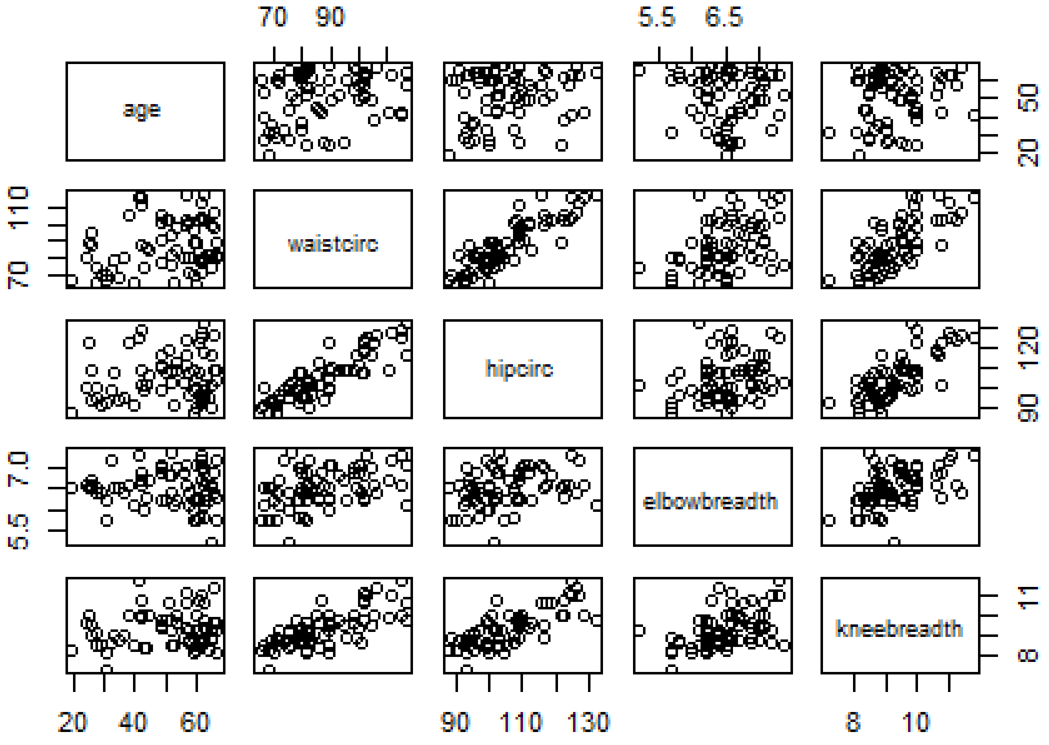


Figure 1.2: Bodyfat Correlation Plot 2



Bodyfat Correlation Plot 1
base R



Bodyfat Correlation Plot 2
base R

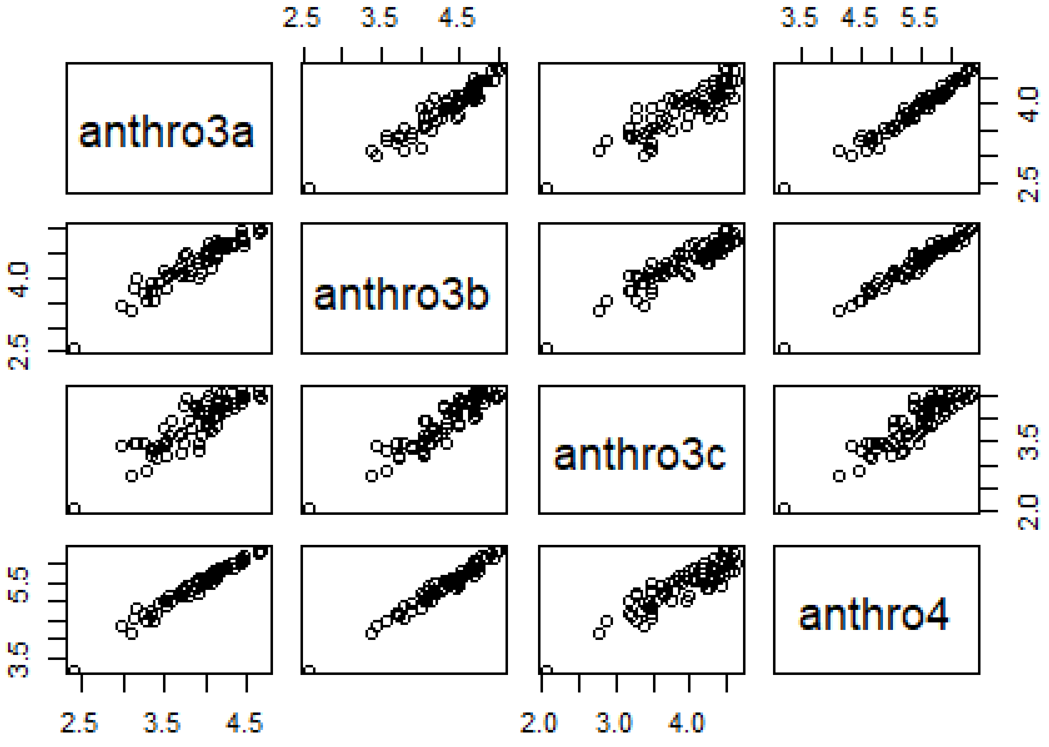


Figure 1.3: Correlation Between Predictors

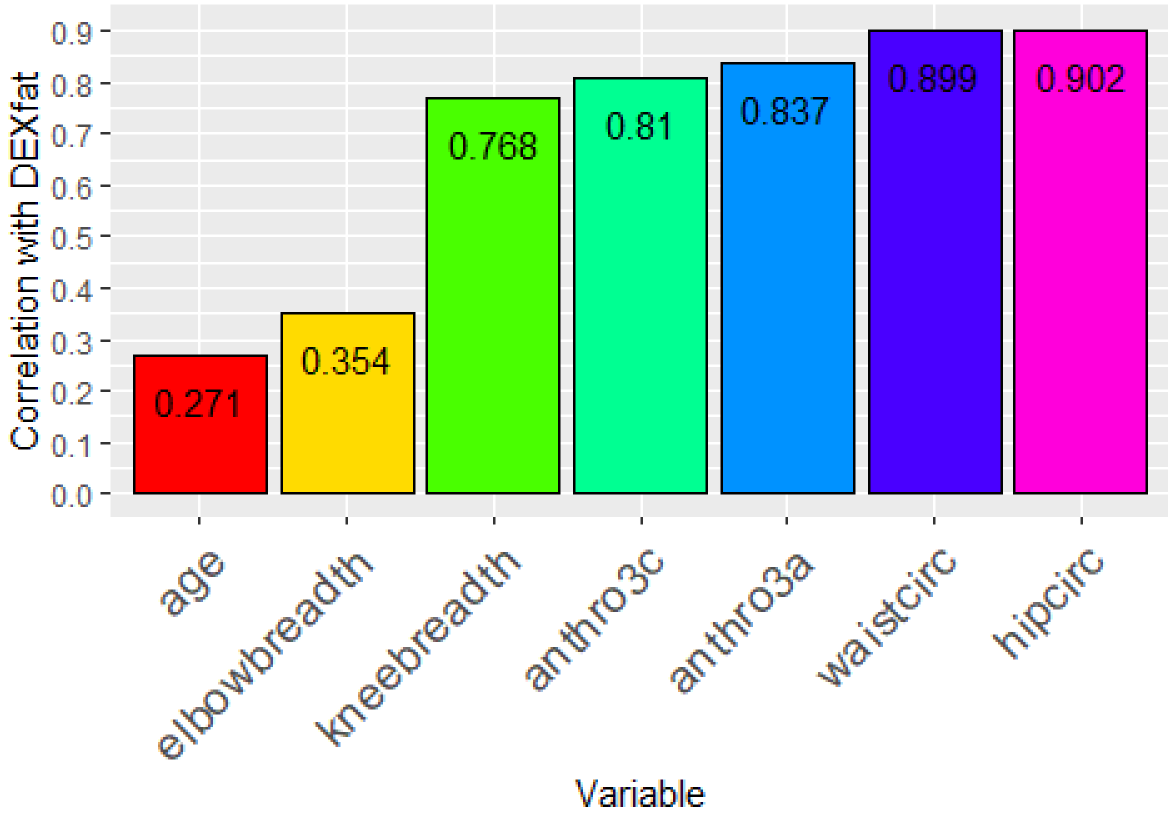
	waistcirc	hipcirc	elbowbreadth	kneebreadth	anthro3a	anthro3b	anthro3c	anthro4
age	0.239	0.180	0.0667	0.128	0.334	0.332	0.281	0.345
waistcirc	0.000	0.871	0.4009	0.732	0.755	0.707	0.740	0.740
hipcirc	0.000	0.000	0.3335	0.759	0.711	0.666	0.689	0.688
elbowbreadth	0.000	0.000	0.0000	0.463	0.325	0.253	0.241	0.294
kneebreadth	0.000	0.000	0.0000	0.000	0.583	0.512	0.538	0.540
anthro3a	0.000	0.000	0.0000	0.000	0.000	0.951	0.884	0.982
anthro3b	0.000	0.000	0.0000	0.000	0.000	0.000	0.933	0.984
anthro3c	0.000	0.000	0.0000	0.000	0.000	0.000	0.000	0.917
anthro4	0.000	0.000	0.0000	0.000	0.000	0.000	0.000	0.000

Figure 1.4: Included Predictor Variables

Variables
age
waistcirc
hipcirc
elbowbreadth
kneebreadth
anthro3a
anthro3c

Correlation with DEXfat by Independent Variable

Figure 1.5



Correlation with DEXfat by Independent Variable-baseR

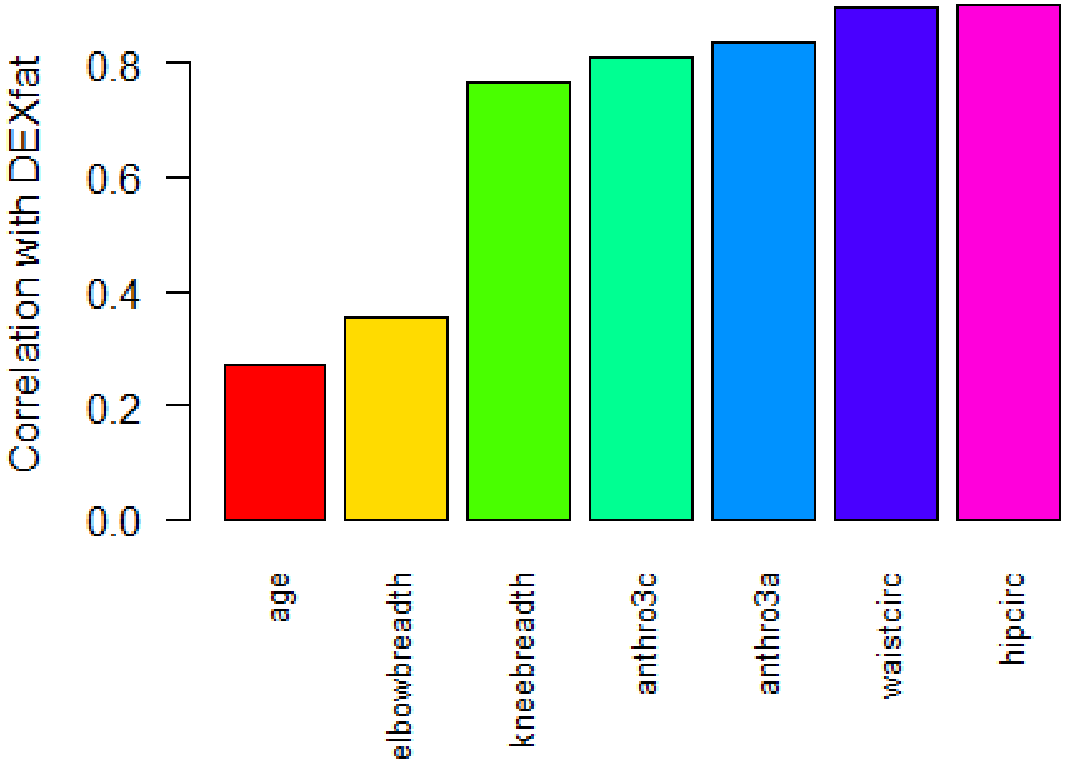


Figure 1.6: Variable Correlation with DEXfat

Independent Variable	Correlation with DEXfat
age	0.2710550
elbowbreadth	0.3535732
kneebreadth	0.7680517
anthro3c	0.8095437
anthro3a	0.8370327
waistcirc	0.8986535
hipcirc	0.9021881

Problem #1, Part B: Fit a generalized additive model assuming normal errors using function **gam**- 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. 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 diagnostic plot. Does it appear that the normality assumption is violated?
- Write a discussion on all of but not limited to the above points.

Results: *Figure 1.7* shows the respective p-values for each variable as it is included in the model provided for the assignment. At an alpha of 0.05, the following variables are statistically significant predictors of 'DEXfat': **waistcirc, hipcirc, kneebreadth, anthro3a**.

Figures 1.8 show that a couple of variables need to be smoothed more than other. These variables are shown in *Figure 1.8(v)* and *Figure 1.8(vii)* and represent **kneebreadth** and **anthro3c**. *Figure 1.8(iii)* also indicates that **hipcirc** may need to be smoothed for model accuracy.

Figure 1.9 shows the GCV (**8.15**), AIC (**344.01**), R-Squared (**.9536**) and degrees of freedom (**20.72**) of the model supplied by this exercise. The R-squared value shows that 95.36% of the variation in 'DEXfat' can be explained by the model as provided.

Finally, we see the summary provided by *gam.check*, as the instructions requested. We see from the plot titled **Resids. vs. linear pred.** that the residuals are relatively normalized around 0. The included histogram shows the same type of distribution. We also see that the **Response vs Fitted Values** plot is linear in nature and appears thata this model is a pretty good predictor - as we saw by looking at R-Squared in *Figure 1.9*.

Figure 1.7: P-Value by Variable

	P-Value
s(age)	0.3130172
s(waistcirc)	0.0010314
s(hipcirc)	0.0000924
s(elbowbreadth)	0.9720289
s(kneebreadth)	0.0000009
s(anthro3a)	0.0004551
s(anthro3c)	0.0822401

Figure 1.8(i)

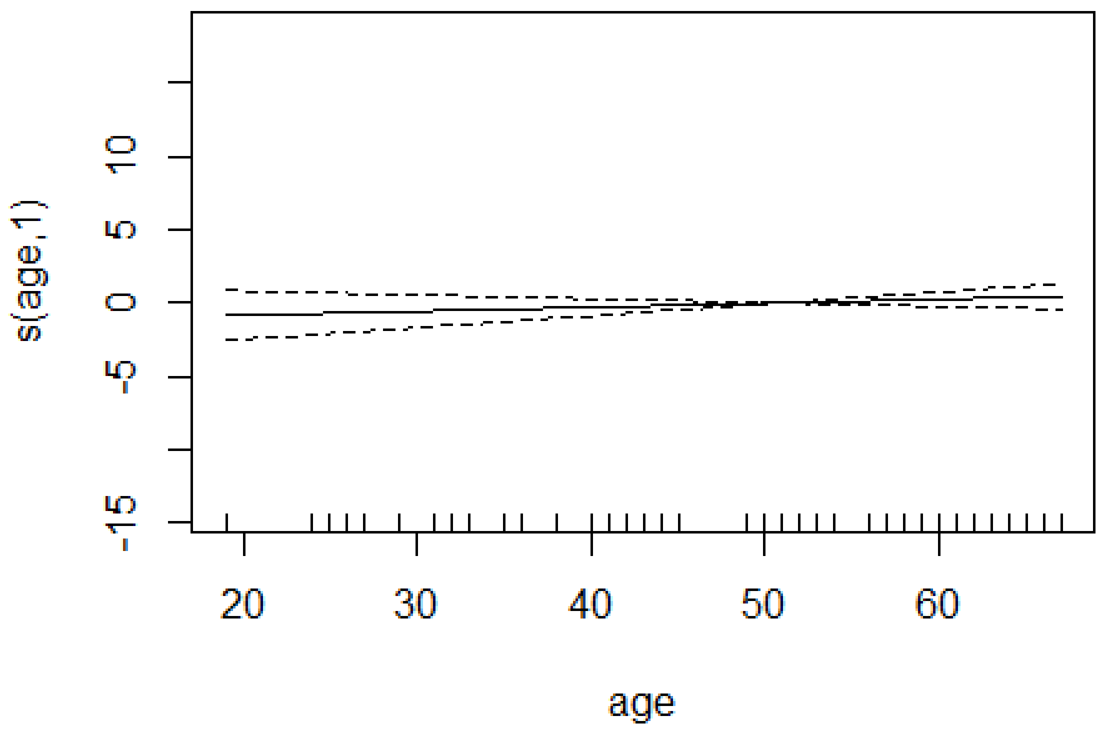


Figure 1.8(ii)

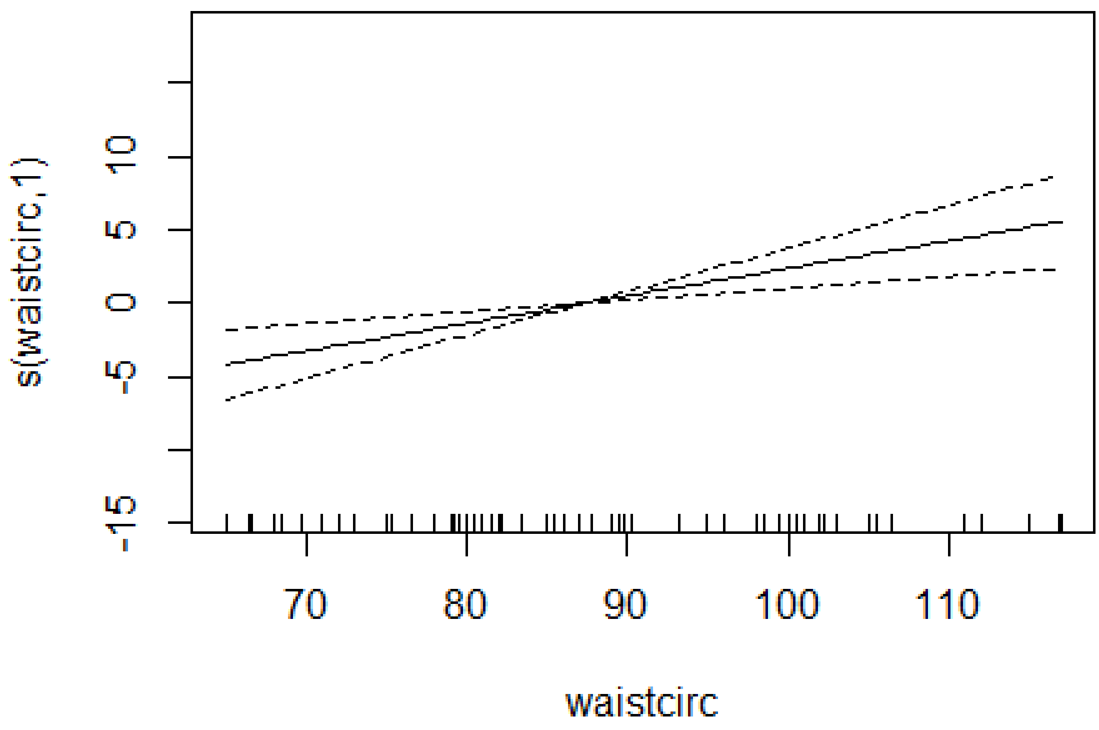


Figure 1.8(iii)

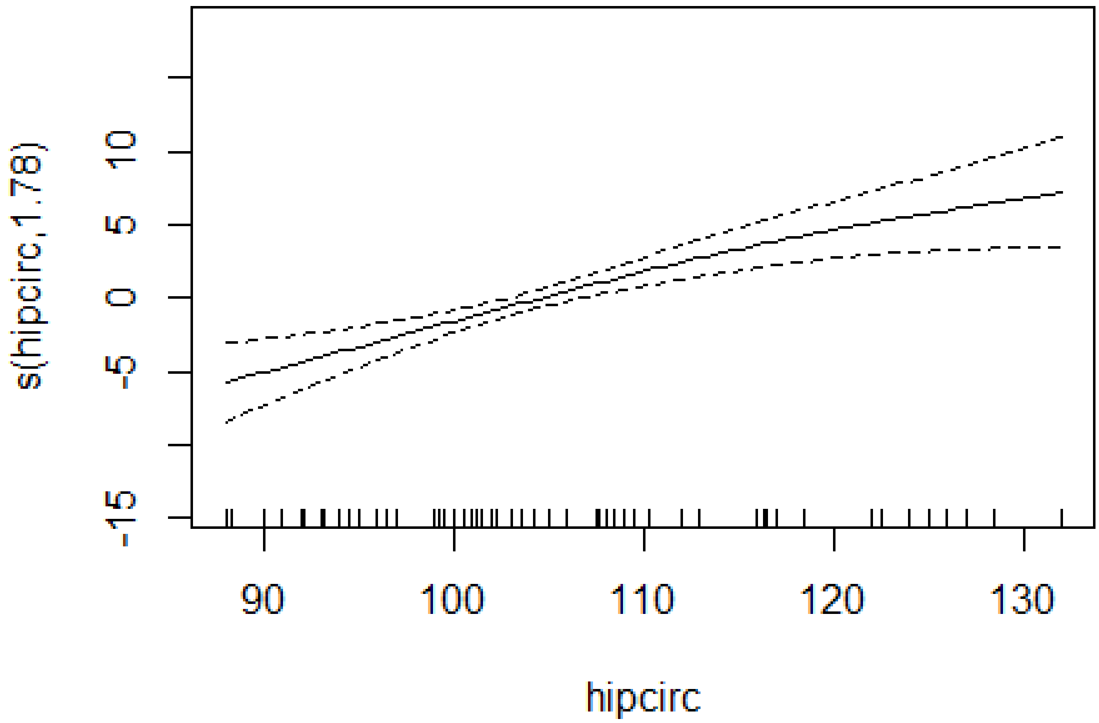


Figure 1.8(iv)

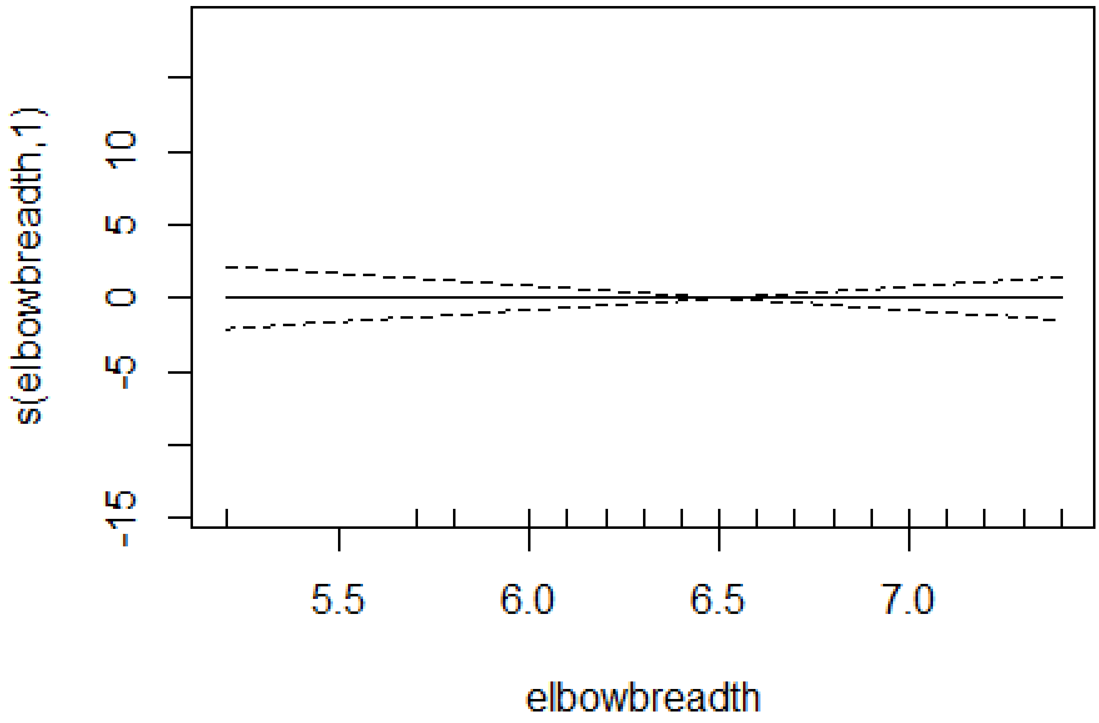


Figure 1.8(v)

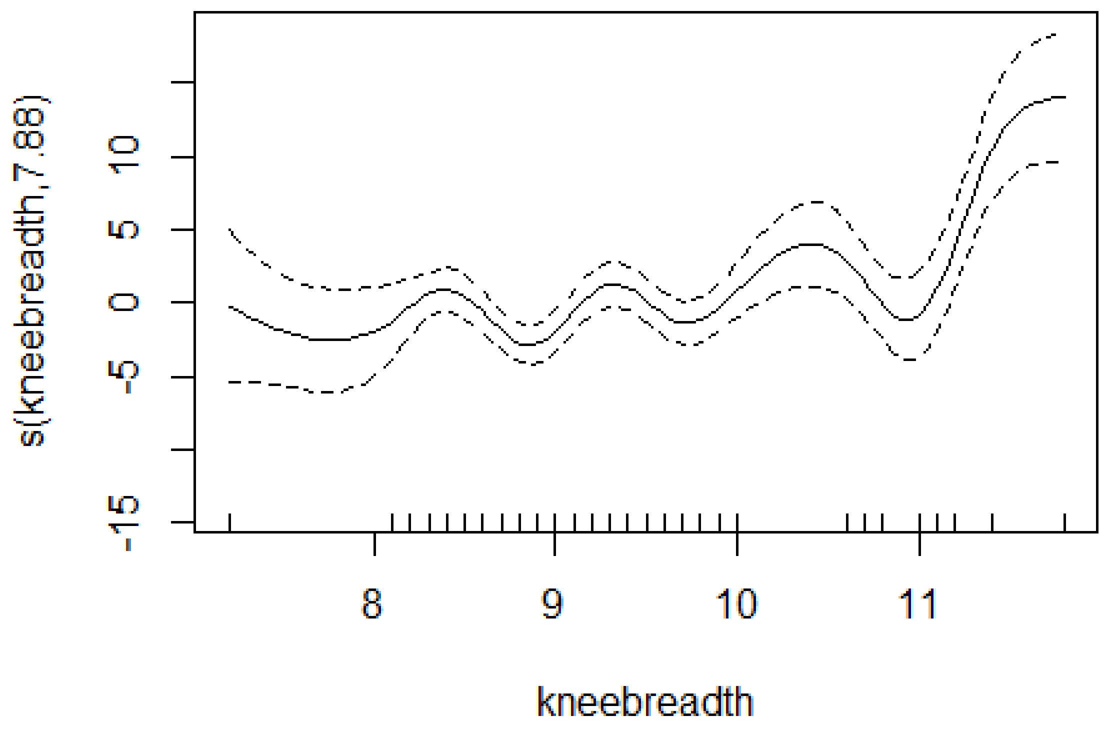


Figure 1.8(vi)

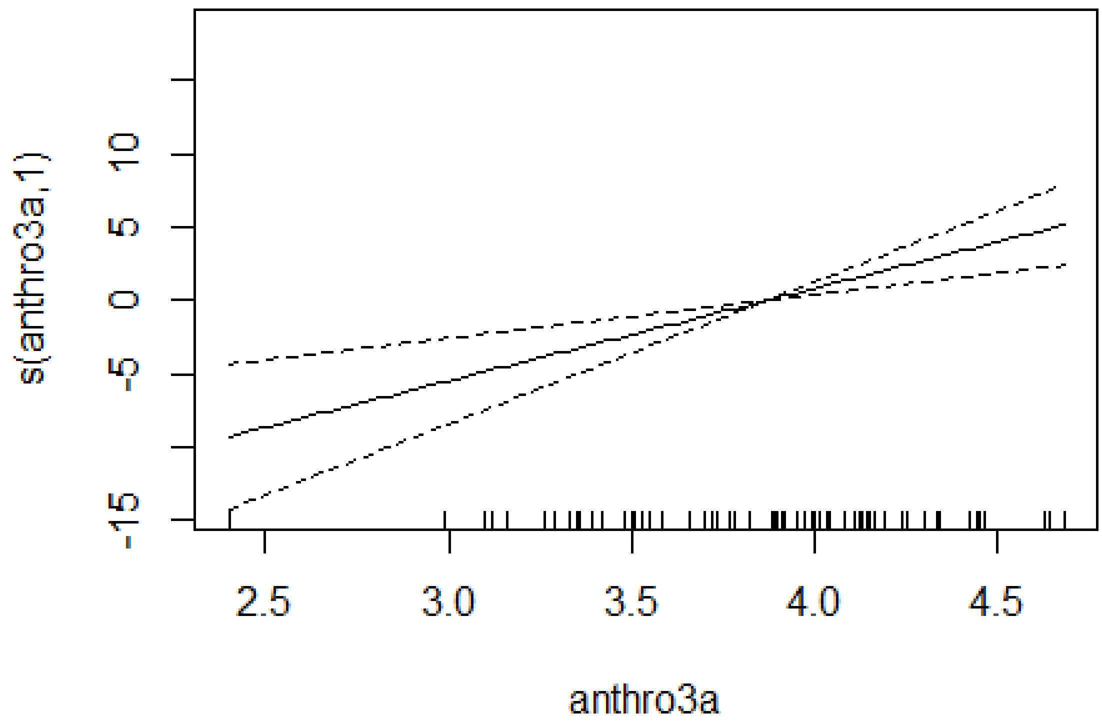


Figure 1.8(vii)

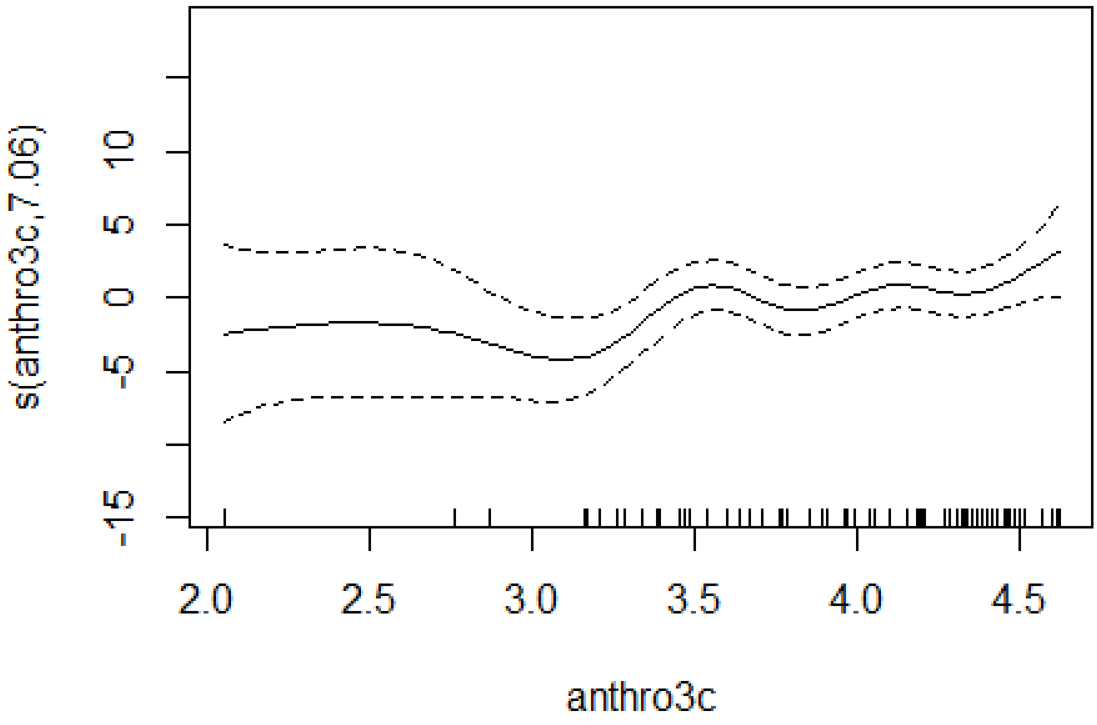
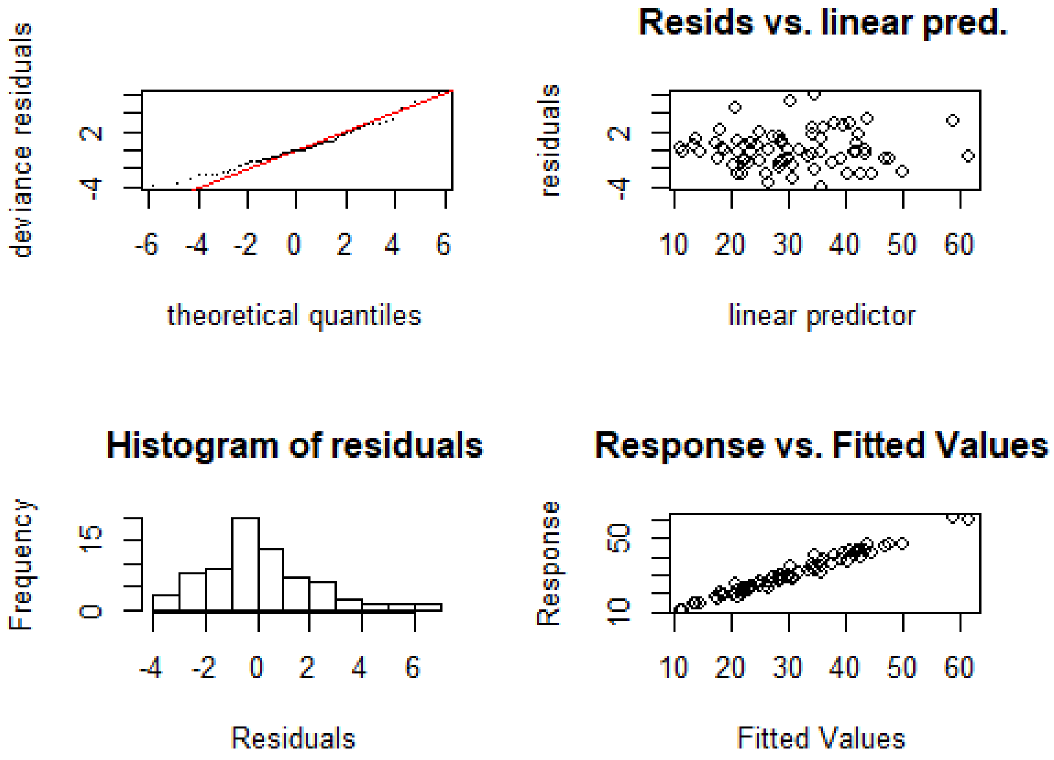


Figure 1.9: Model Statistics Summary

GCV	AIC	R2	DF
8.147249	344.0106	0.9536278	20.72114



```
## 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 =  63 / 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    0.06 .
## s(waistcirc)    9.00 1.00    0.94    0.26
## s(hipcirc)      9.00 1.78    1.02    0.54
## s(elbowbreadth) 9.00 1.00    0.81    0.03 *
## s(kneebreadth)  9.00 7.88    1.08    0.71
## s(anthro3a)     9.00 1.00    1.09    0.69
## s(anthro3c)     9.00 7.06    0.89    0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem #1, Part C: Now remove insignificant variables and remove smoothing for some variables. Report summary, plot, GCV, AIC, adj-R2. (Fit the following model as well as another one you come up with on your own, justifying the variables and smoothing you use).

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

Results: *Figures 1.10* show the plots of the smoothed predictors from the supplied model, per the instructions. We see, similar to the figures above *Figure 1.8(iii), 1.8(v), and 1.8(vii)* above, the need for smoothing of 'hipcirc', 'kneebreadth' and 'anthro3c'.

Per the instructions, a summary of this model is included. The biggest thing I notice from the summary is the statistically significant prediction values for each predictor. All are significant at an alpha of 0.05.

Figures 1.11 show the plots of the two smoothed variables from my alternative model, 'hipcirc' and 'kneebreadth'. For my alternative model, I've dropped 'anthro3c' as a predictor due to it's relatively less significant impact on 'DEXfat' when compared to the other predictors.

Per the homework instructions, a summary of my alternative model is included. This time, we see that all predictors are extremely significant in the prediction of 'DEXfat'.

Figure 1.12 summarizes the comparison statistics from the two models. As we can see from the GCV, AIC, and R-squared, the supplied model that includes 's(anthro3c)' is superior to my alternative model that dropped this predictor. This is to be expected since, although 'anthro3c' was less statistically significant than the other predictors in the supplied model, it was still significant at a level < 0.05.

Figure 1.10(i)

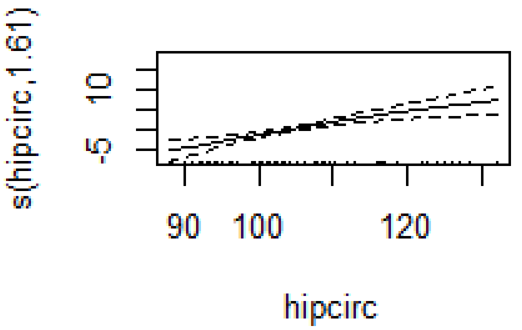


Figure 1.10(ii)

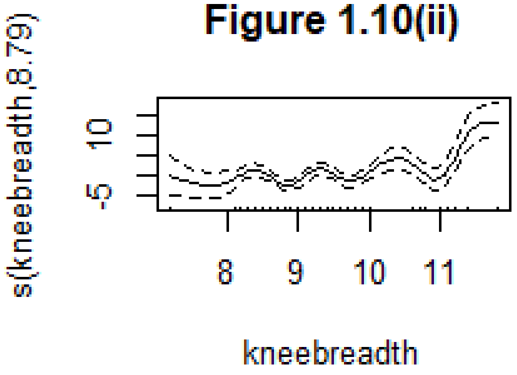
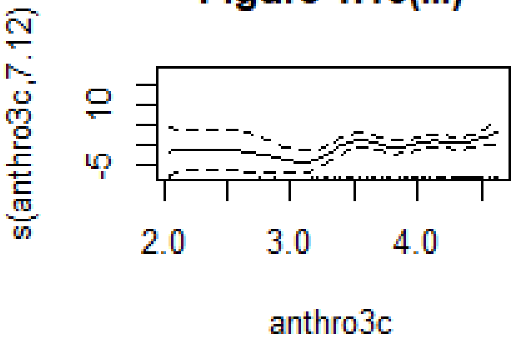


Figure 1.10(iii)



```
## 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
```

Figure 1.11(i)

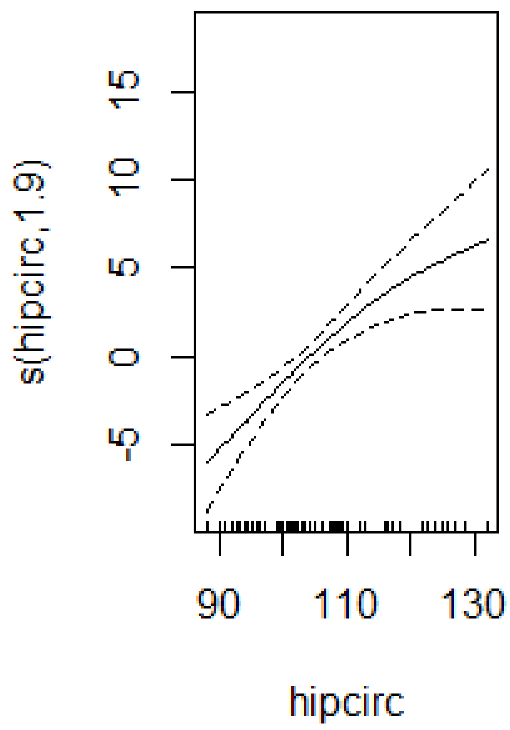
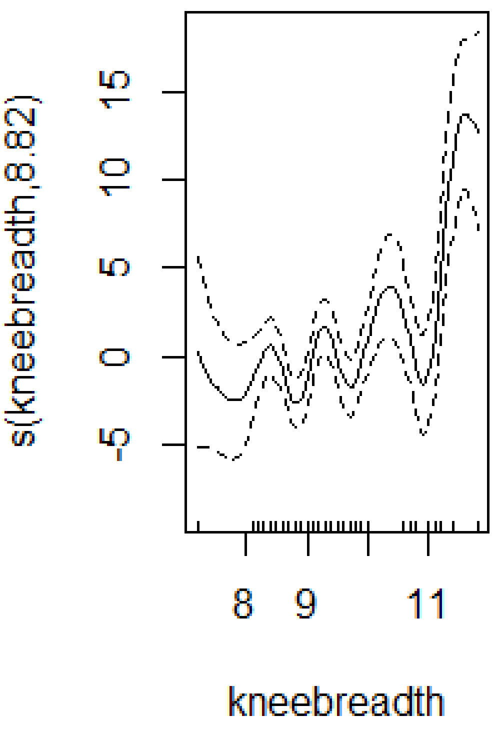


Figure 1.11(ii)



```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## DEXfat ~ waistcirc + s(hipcirc) + s(kneebreadth) + anthro3a
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.35858    5.11469  -3.785  0.00037 ***
## waistcirc     0.24140    0.05268   4.583 2.53e-05 ***
## anthro3a      7.50726    1.09209   6.874 5.04e-09 ***
## ---
## 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)     1.899  2.397 8.899 0.000218 ***
## s(kneebreadth) 8.825  8.986 5.527 7.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.944   Deviance explained = 95.4%
## GCV = 8.5004  Scale est. = 6.8573    n = 71
```

Figure 1.12: 2nd Model Statistics Summary

GCV	AIC	R2
7.946447	343.2562	0.9536683
8.500360	352.3836	0.9437664

Problem #1, Part 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, Adj-R2, residual plots, etc. to compare models).

Results: *Figure 1.13* gives us our model statistics using the supplied model from *part C*, as well as my alternative model from *part C*, with the log-transformed response model from this part of the exercise. Per the question, we are comparing the three models, but according to Burnham and Anderson (*Model Selection and Multi-Model Inference*, 2004) comparing the 3 is not possible since we've transformed the response in one of the models.

We see that the generalized cross-validation (GCV) score of the GAM fitted model, with log transformed response, is very low compared to the previous 2 models. The AIC is lower - actually negative - in the log transformed response model. This is the first time I've seen a negative AIC but evidently AIC should not be exclusively non-negative.

The r-squared of the supplied model from *part C* is superior indicating that this model explains the highest percentage of variation in 'DEXfat' of the 3 models.

Figure 1.13: Log Model Statistics Summary

	GCV	AIC	R2
Supplied Model	7.9464465	343.2562	0.9536683
Alternative Model	8.5003603	352.3836	0.9437664
Log Response Model	0.0088137	-136.4700	0.9522733

Problem #1, Part E: Fit generalized additive model that underwent AIC-based variable selection (fitted using function **gamboost()**). 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)]

Results: The variable dropped by the AIC-based variable selection, fitted using *gamboost* is **age**, as we can see from *Figure 1.14*. This is somewhat predictable given that it has the lowest correlation with **DEXfat** as I showed in *Figure 1.6*.

Figure 1.14: Variable Removed by GAMBoost

variable
age

Problem #2: Fit a logistic additive model to the glaucoma data. (Here use family = "binomial"). Which covariates should enter 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 model.)

Results: First I fit a logistic additive model, using gamboost due to the number of covariates. *Figure 2.1* shows the variables that have been included in the model based on gamboost.

Figure 2.2 shows the selection probabilities of each variable. Their ranking indicates the influence each variable has on the probability of suffering from Glaucoma.

Figure 2.1: Variables Kept by GAMBoost

variables
as
abrs
hic
mhcg
mhcn
mhci
phcg
phcn
phci
hvc
vass
vars
vari
mdn
mdi
tms
tmi
mv

Figure 2.2: Variable Importance

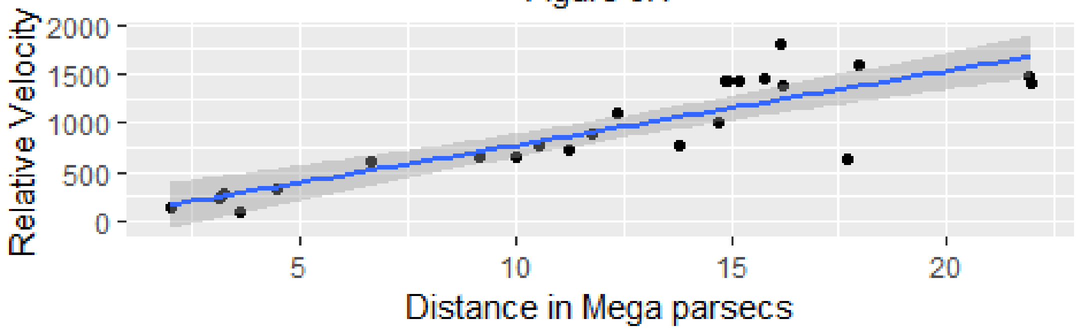
Variables	Selection Probs
tmi	0.17
mhcg	0.11
vars	0.11
mhci	0.10
hvc	0.08
vass	0.08
as	0.07
vari	0.06
mv	0.04
abrs	0.03
mgcn	0.03
phcn	0.03
mdn	0.03
phci	0.02
hic	0.01
phcg	0.01
mdi	0.01
tms	0.01

Problem #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).

Results: *Figure 3.1* and *Figure 3.2* show the different types of scatterplot smoothers similar to what is using on the men1500m data in chapter 10. I've included analogous base R plots.

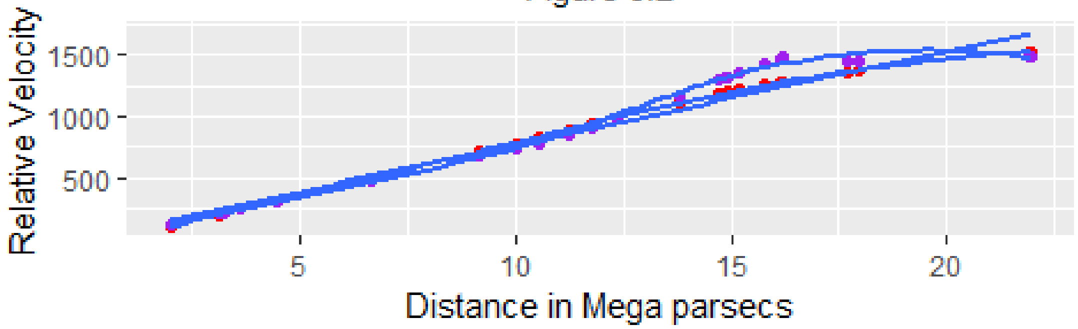
Linear Model Scatterplot

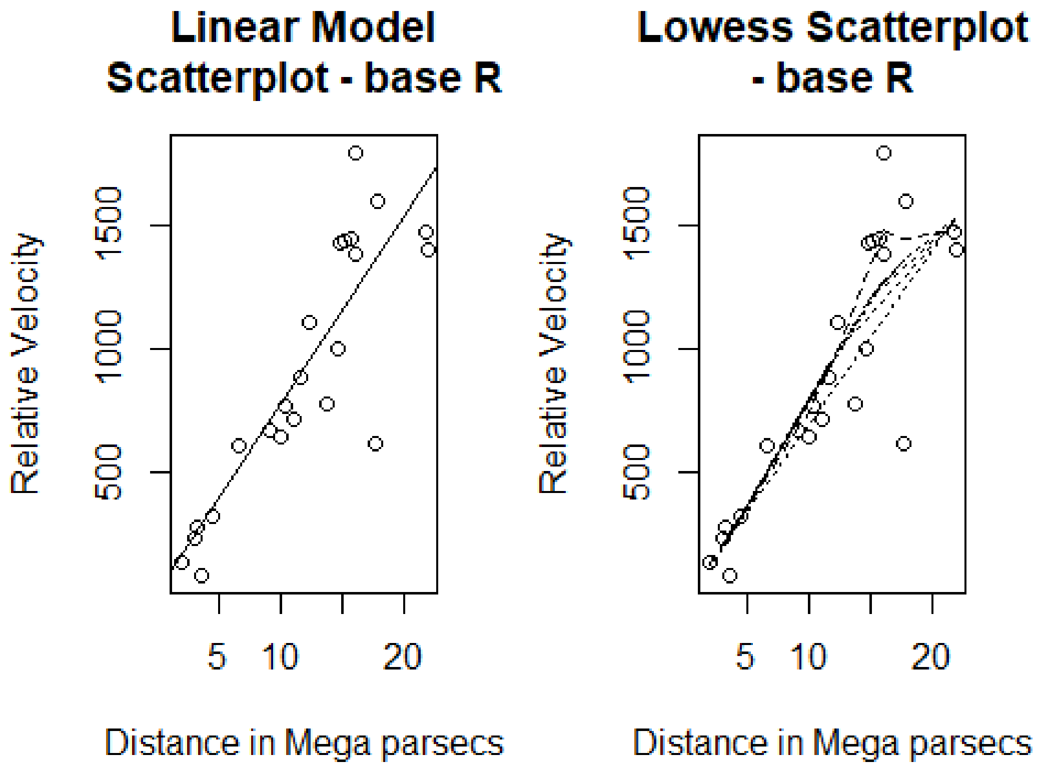
Figure 3.1



Hubble Scatterplot with Predictors

Figure 3.2


Relative Velocity
Distance in Mega parsecs


Linear Model Scatterplot - base R
Relative Velocity
Distance in Mega parsecs
Lowess Scatterplot - base R
Relative Velocity
Distance in Mega parsecs