GAMs for Hirsutism data

Gregoire Gasparini, Aurora Hofman, Beatriu Tort

30 de marzo de 2020

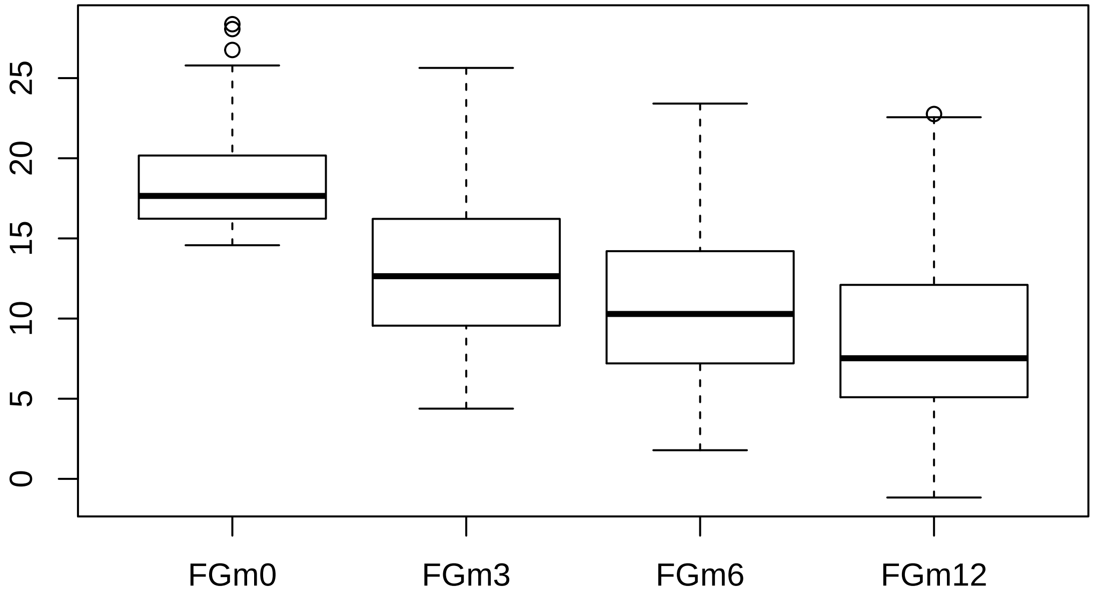
Hirsutism dataset

```
hirs <- read.table("hirsutism.dat",header=T, sep="\t",fill=TRUE)
Treatment<- hirs$Treatment <- as.factor(hirs$Treatment)
summary(hirs)
```

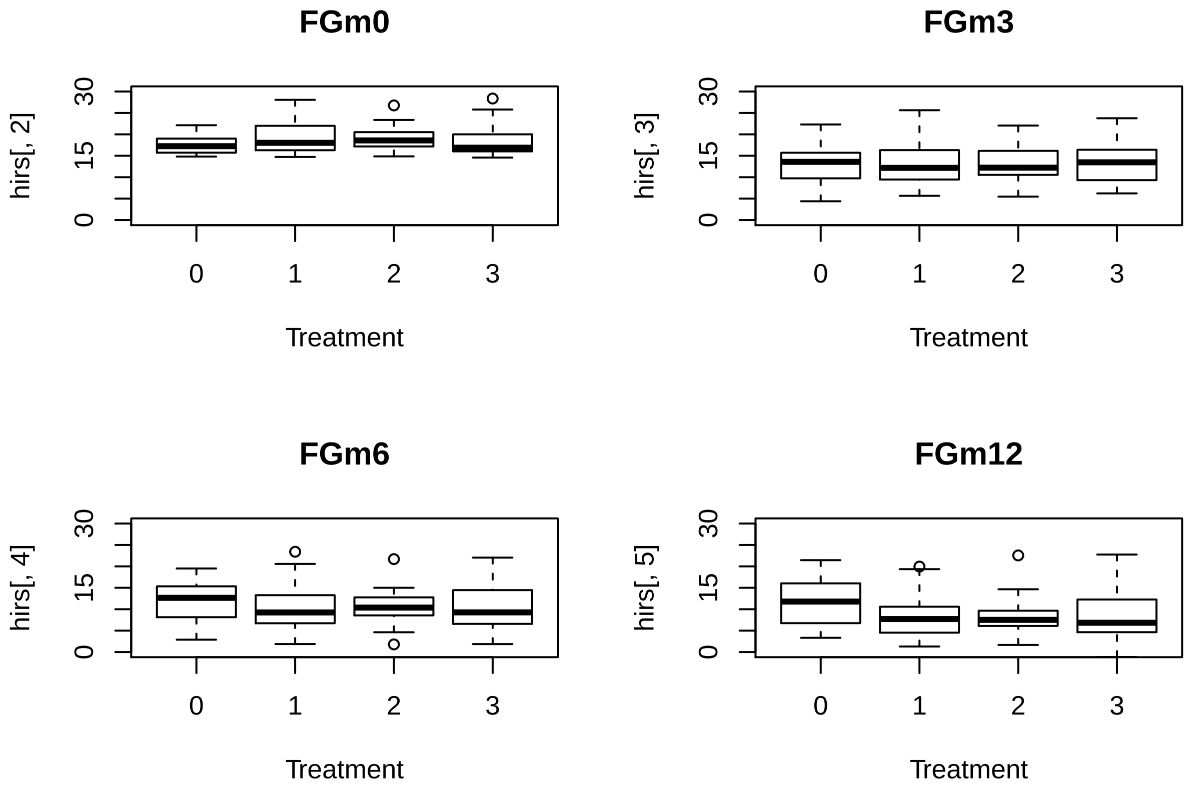
```
##  Treatment      FGm0            FGm3             FGm6             FGm12
##  0:23      Min.   :14.57   Min.   : 4.381   Min.   : 1.786   Min.   :-1.163
##  1:26      1st Qu.:16.23   1st Qu.: 9.557   1st Qu.: 7.202   1st Qu.: 5.093
##  2:24      Median :17.65   Median :12.643   Median :10.286   Median : 7.524
##  3:26      Mean   :18.57   Mean   :13.084   Mean   :10.853   Mean   : 8.911
##            3rd Qu.:20.17   3rd Qu.:16.219   3rd Qu.:14.204   3rd Qu.:12.101
##            Max.   :28.36   Max.   :25.637   Max.   :23.411   Max.   :22.759
##
##     SysPres         DiaPres          weight           height
##  Min.   : 88.0   Min.   :46.00   Min.   : 41.00   Min.   :1.480
##  1st Qu.:110.0   1st Qu.:65.00   1st Qu.: 57.00   1st Qu.:1.580
##  Median :115.0   Median :70.00   Median : 64.00   Median :1.610
##  Mean   :115.9   Mean   :70.04   Mean   : 68.06   Mean   :1.613
##  3rd Qu.:120.0   3rd Qu.:75.00   3rd Qu.: 74.50   3rd Qu.:1.650
##  Max.   :162.0   Max.   :95.00   Max.   :113.00   Max.   :1.800
##  NA's   :8       NA's   :8       NA's   :8        NA's   :8
```

Boxplots to get an overview of the data.

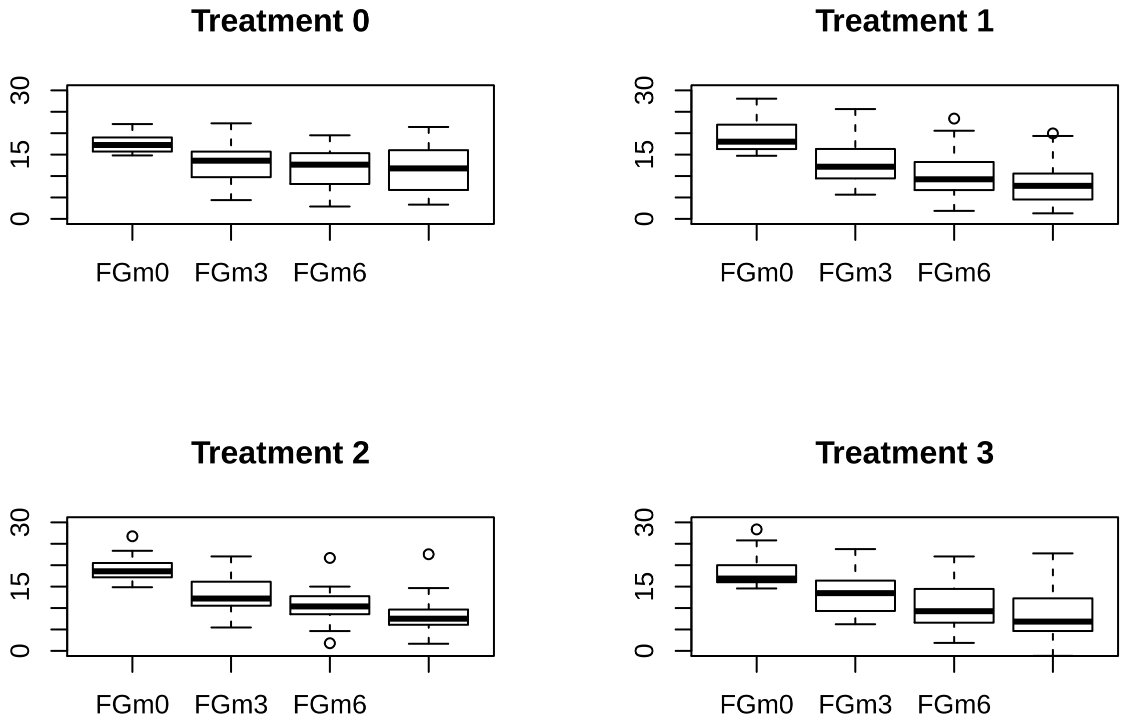
```
boxplot(hirs[,2:5])
```



```
par(mfrow=c(2,2))
boxplot(hirs[,2]~hirs$Treatment,ylim=c(0,30), main=names(hirs)[2], xlab="Treatment")
boxplot(hirs[,3]~hirs$Treatment,ylim=c(0,30), main=names(hirs)[3], xlab="Treatment")
boxplot(hirs[,4]~hirs$Treatment,ylim=c(0,30), main=names(hirs)[4], xlab="Treatment")
boxplot(hirs[,5]~hirs$Treatment,ylim=c(0,30), main=names(hirs)[5], xlab="Treatment")
```

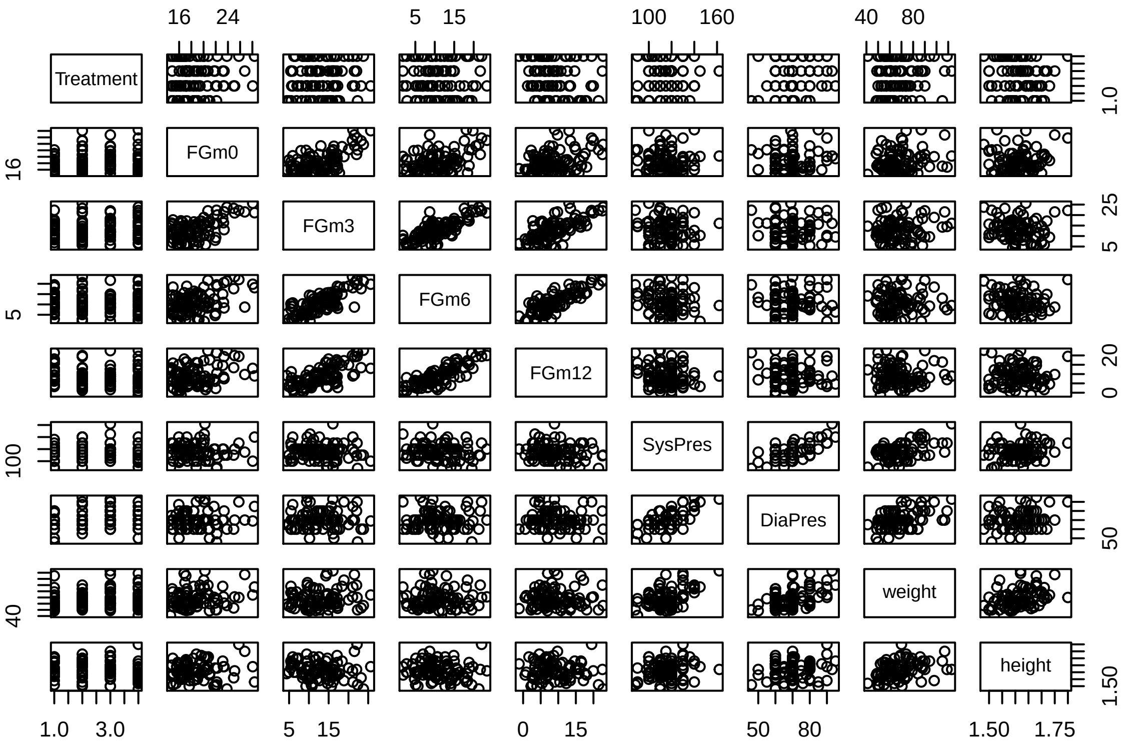


```
par(mfrow=c(1,1))
par(mfrow=c(2,2))
boxplot(hirs[hirs$Treatment==0,2:5],ylim=c(0,30), main="Treatment 0")
boxplot(hirs[hirs$Treatment==1,2:5],ylim=c(0,30), main="Treatment 1")
boxplot(hirs[hirs$Treatment==2,2:5],ylim=c(0,30), main="Treatment 2")
boxplot(hirs[hirs$Treatment==3,2:5],ylim=c(0,30), main="Treatment 3")
```



```
par(mfrow=c(1,1))

pairs(hirs)
```



1st GAM model: linear model through GAM (FGm12 ~ Treatment + FGm0 + SysPres + DiaPres + weight + height)

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-31. For overview type 'help("mgcv-package")'.
```

```
gam1 <- gam(FGm12 ~ Treatment + FGm0 + SysPres + DiaPres + weight + height, data = hirs)
summary(gam1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ Treatment + FGm0 + SysPres + DiaPres + weight + height
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.49686   14.85245   1.313 0.192945
## Treatment1  -4.33022    1.48110  -2.924 0.004471 **
## Treatment2  -4.31441    1.49589  -2.884 0.005012 **
## Treatment3  -3.94666    1.44364  -2.734 0.007668 **
## FGm0         0.59983    0.16862   3.557 0.000626 ***
## SysPres     -0.07570    0.05194  -1.458 0.148787
## DiaPres      0.03525    0.07115   0.495 0.621652
## weight       0.02768    0.04425   0.626 0.533308
## height      -8.71024    9.08570  -0.959 0.340540
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =   0.17   Deviance explained = 24.4%
## GCV = 25.139  Scale est. = 22.653    n = 91
```

2nd GAM model: smooth model through GAM (FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) + s(DiaPres) + s(weight) + s(height))

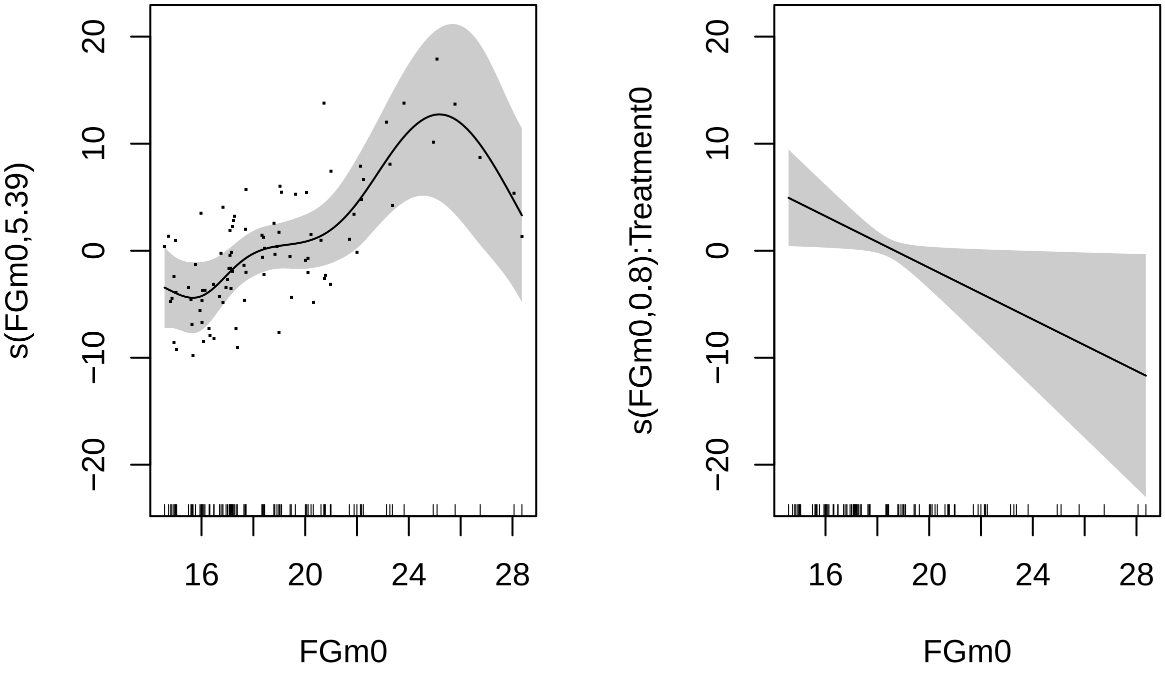
We fit a full model with all the possible explanitory variables.

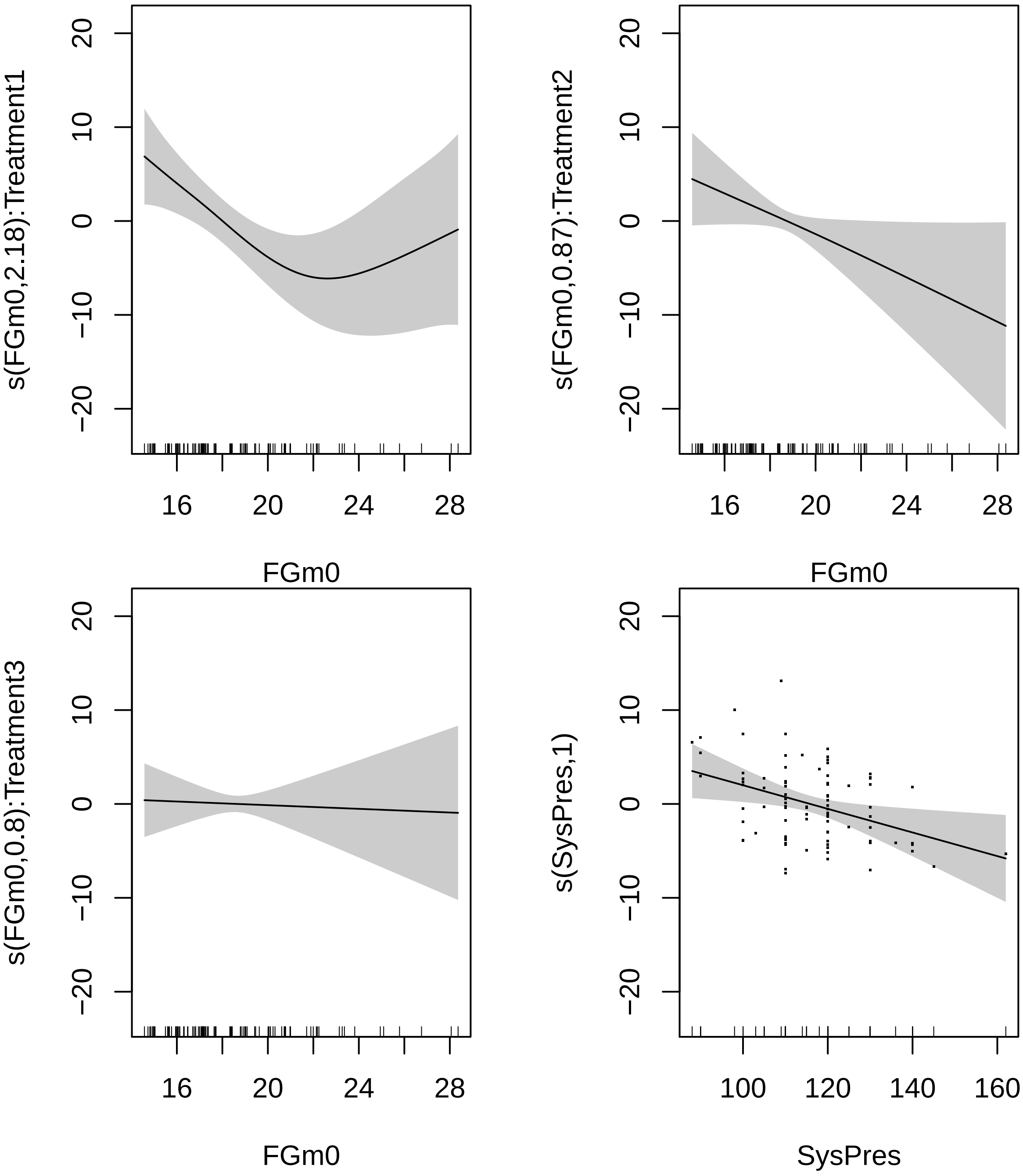
```
gam2 <- gam(FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) + s(DiaPres) + s(weight)
summary(gam2)
```

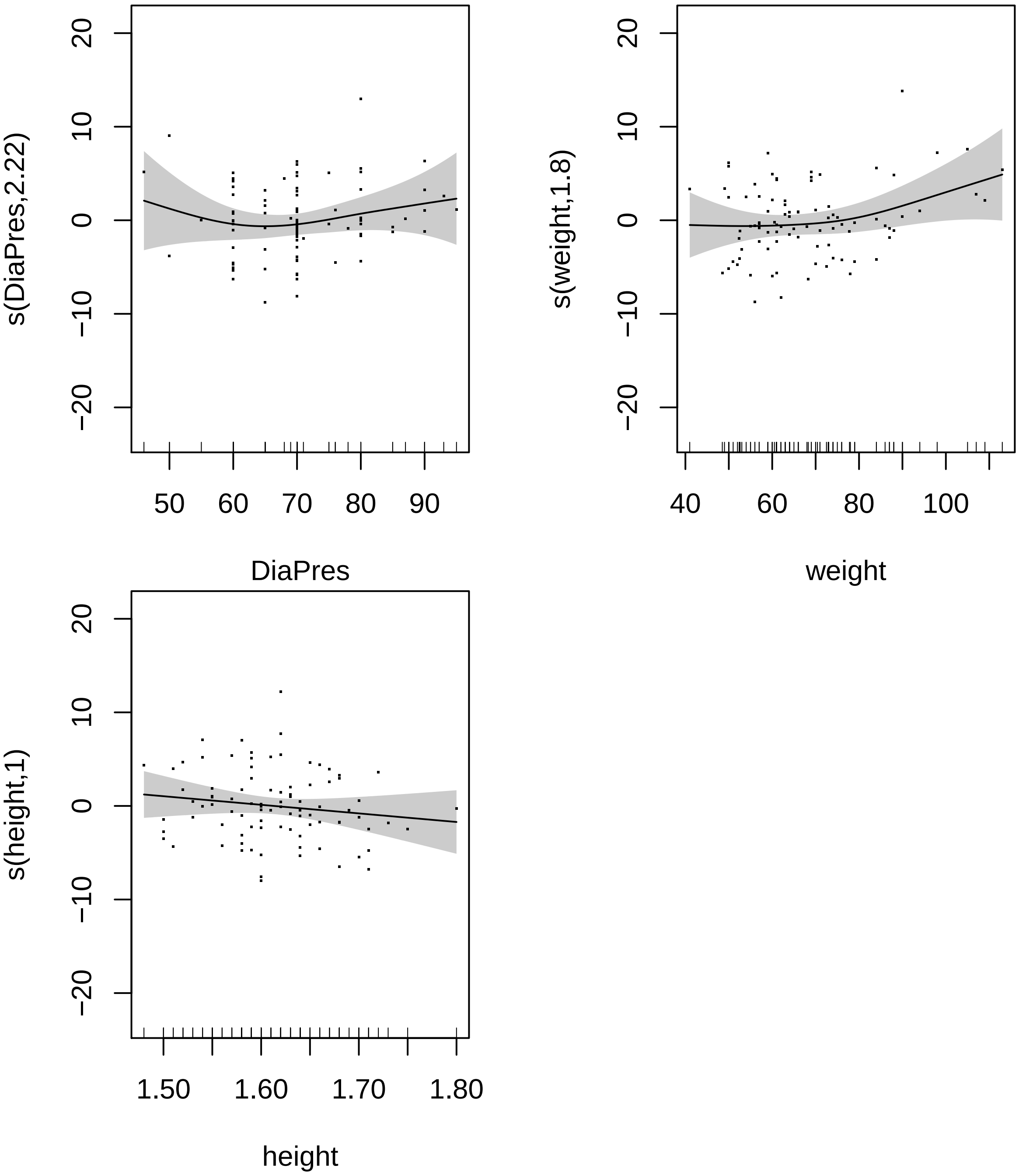
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) +
##     s(DiaPres) + s(weight) + s(height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.765      1.075  10.941  < 2e-16 ***
## Treatment1    -3.888      1.472  -2.641  0.01017 *
## Treatment2    -3.837      1.426  -2.691  0.00888 **
## Treatment3    -3.256      1.440  -2.261  0.02686 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df     F p-value
## s(FGm0)            5.3929  6.464 2.472  0.0384 *
## s(FGm0):Treatment0 0.8000  0.800 5.569  0.0383 *
## s(FGm0):Treatment1 2.1794  2.762 3.449  0.0348 *
## s(FGm0):Treatment2 0.8712  0.933 4.289  0.0492 *
## s(FGm0):Treatment3 0.8000  0.800 0.053  0.8372
## s(SysPres)         1.0000  1.000 6.514  0.0128 *
## s(DiaPres)         2.2156  2.783 1.262  0.3929
## s(weight)          1.8031  2.261 1.991  0.1345
## s(height)          1.0000  1.000 1.093  0.2992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 84/85
## R-sq.(adj) =  0.368   Deviance explained = 50.2%
## GCV = 22.134  Scale est. = 17.254    n = 91
```

From this model we can see that some variables potentially can be removed as they have a p-value that implies they are not significant. Also the variable SysPres which is significant has a edf of 1 so it can be replaced by linear terms.

```
plot(gam2, residuals = TRUE, shade=TRUE, seWithMean=TRUE, pages = 7)
```




s(FGm0,2.18):Treatment1
s(FGm0,0.87):Treatment2
s(FGm0,0.8):Treatment3
s(SysPres,1)
20
10
0
−10
−20
16
20
24
28
FGm0
FGm0
FGm0
100
120
140
160
SysPres



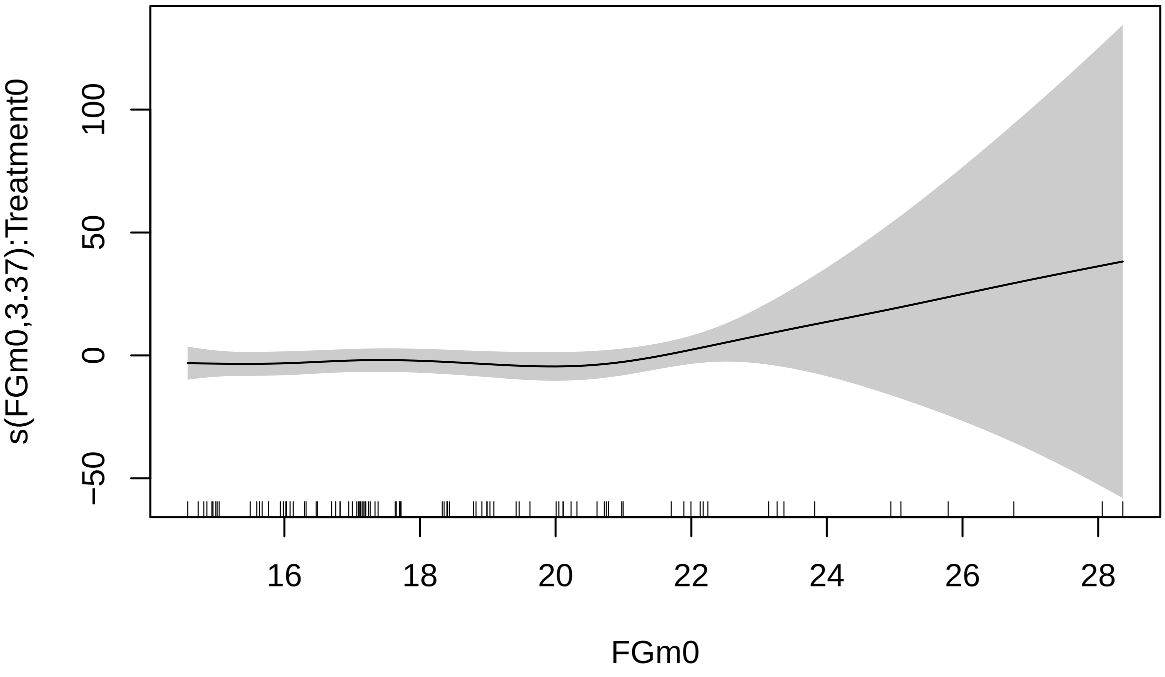
From this model we can see that some variables can be removed as they have a p-value that implies they are not significant. Also the variable SysPres which is significant has a edf of 1 so it can be replaced by linear terms, this can also be seen from the plot. This leads us to the following model where s(DiaPres) + s(weight) + s(height) are removed and s(SysPres) is replaced by a linear term leading to a semiparametric model.

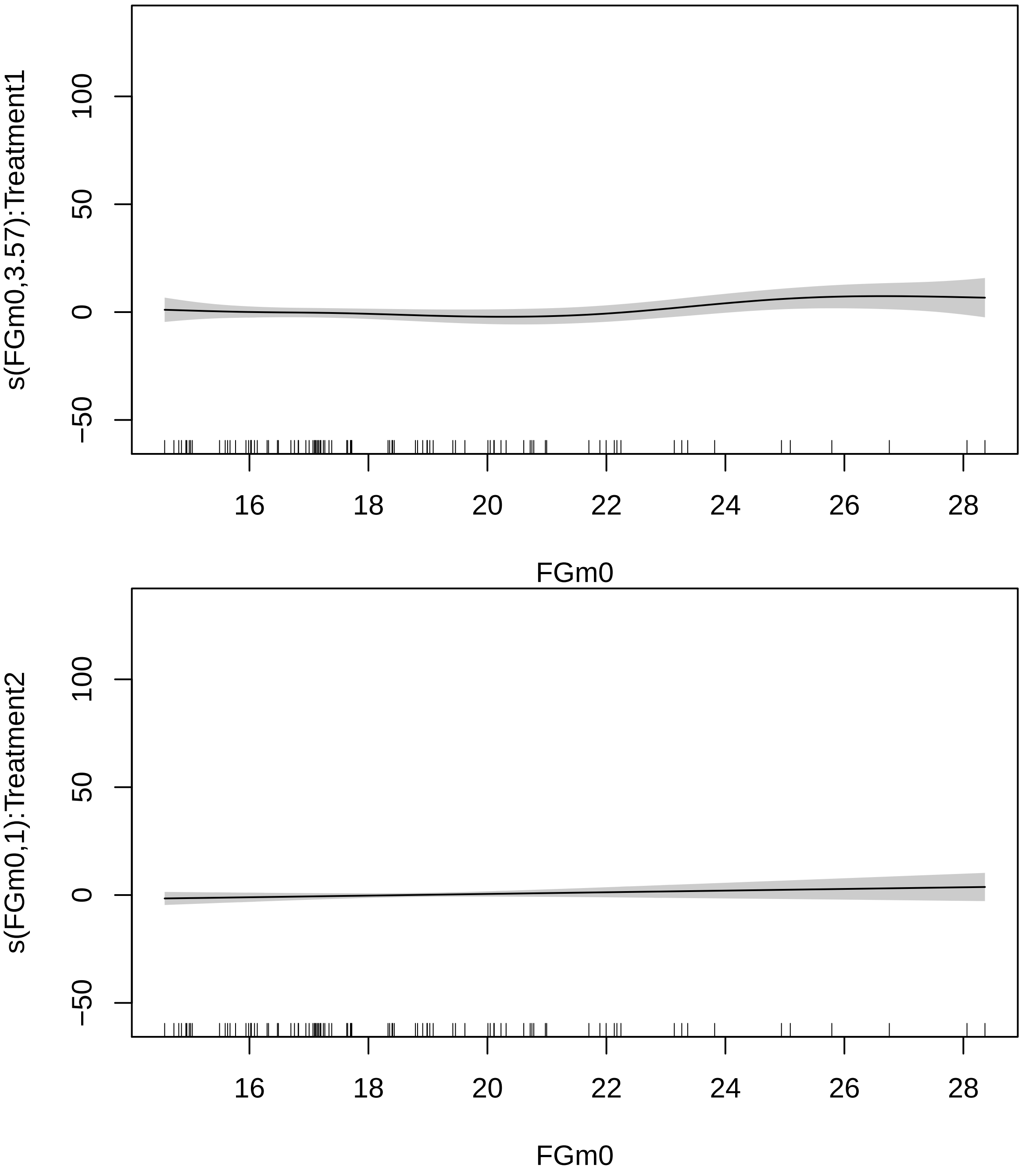
3rd GAM model: smooth model through GAM (FGm12 ~ Treatment + s(FGm0, by = Treatment) + SysPres)

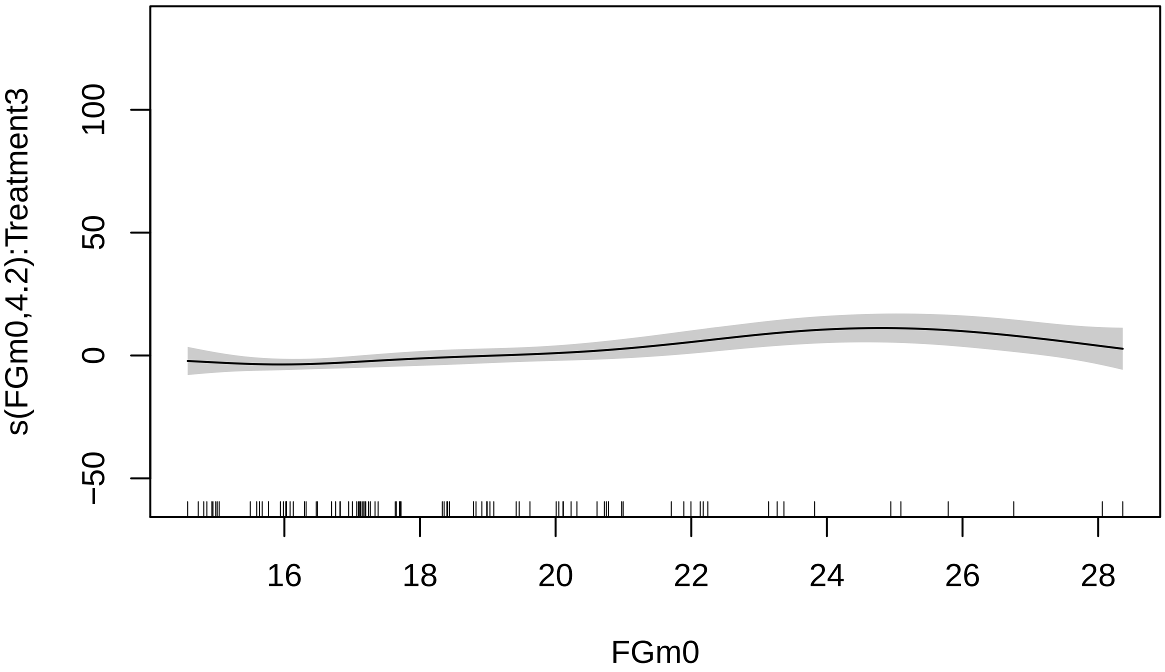
```
gam3 <- gam(FGm12 ~ Treatment + s(FGm0, by = Treatment) + SysPres, data = hirs)
summary(gam3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ Treatment + s(FGm0, by = Treatment) + SysPres
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.17495    5.04950   3.995 0.000151 ***
## Treatment1  -6.71549    2.87391  -2.337 0.022169 *
## Treatment2  -6.10814    2.87125  -2.127 0.036730 *
## Treatment3  -5.29522    2.82885  -1.872 0.065185 .
## SysPres     -0.05156    0.04056  -1.271 0.207667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                       edf Ref.df     F p-value
## s(FGm0):Treatment0 3.367  4.085 0.563 0.75468
## s(FGm0):Treatment1 3.567  4.367 1.762 0.11859
## s(FGm0):Treatment2 1.000  1.000 1.288 0.25998
## s(FGm0):Treatment3 4.203  5.145 4.371 0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.301   Deviance explained = 42.6%
## GCV = 23.523  Scale est. = 19.093    n = 91
```

```r
plot(gam3, residuals = TRUE, shade=TRUE, seWithMean=TRUE)
```

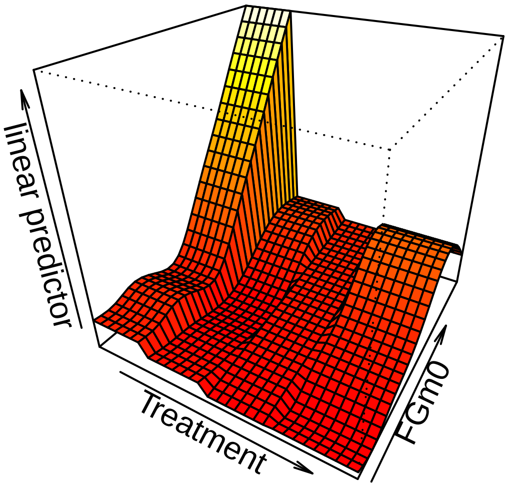



s(FGm0,3.57):Treatment1
100
50
0
−50
16
18
20
22
24
26
28
FGm0
s(FGm0,1):Treatment2
100
50
0
−50
16
18
20
22
24
26
28
FGm0

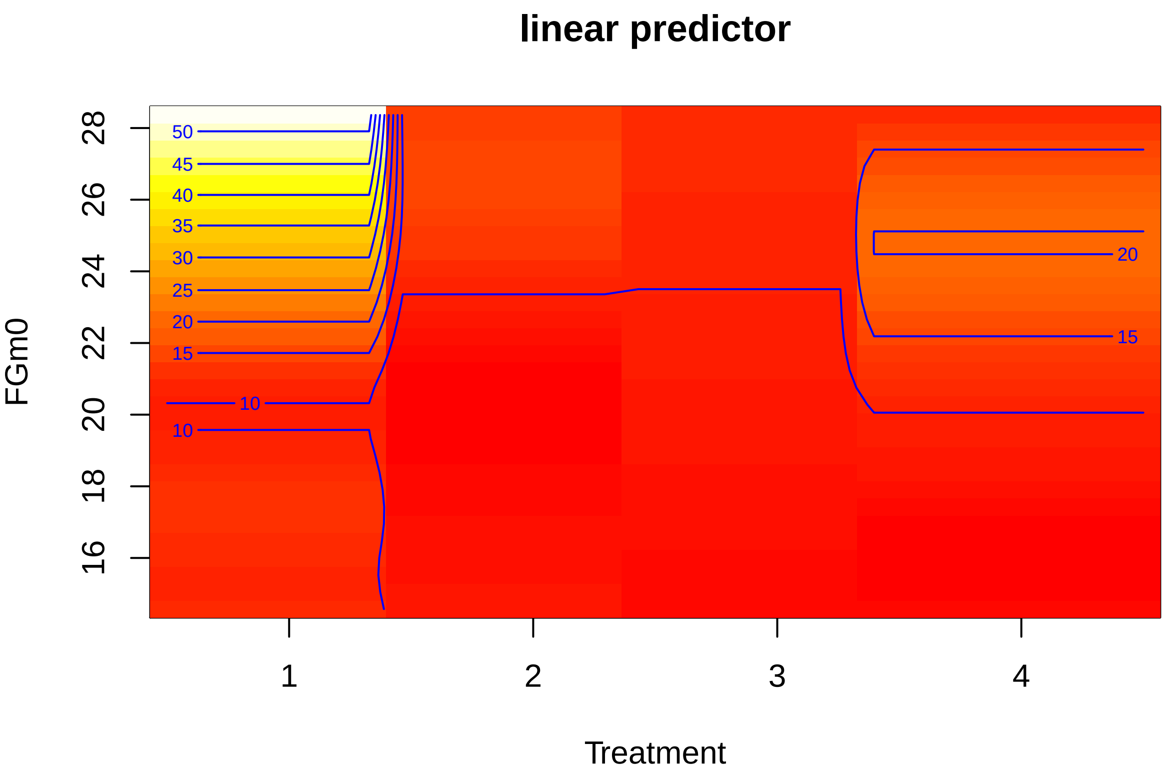


Visualization of the joint effects of variables:

```r
vis.gam(gam3, view=c("Treatment","FGm0"), plot.type = "persp", theta=30, phi=30)
```



```r
vis.gam(gam3, view=c("Treatment","FGm0"), plot.type = "contour")
```



```
anova(gam3,gam2,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ Treatment + s(FGm0, by = Treatment) + SysPres
## Model 2: FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) +
##     s(DiaPres) + s(weight) + s(height)
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    71.403     1410.2
## 2    68.197     1224.0 3.2063   186.26 3.3668 0.02109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the reduced model is rejected for the full model.

4rth GAM model: smooth model through GAM (FGm12 ~ s(FGm0, by = Treatment) + Treatment)

```
gam4 <- gam(FGm12 ~ Treatment + s(FGm0, by = Treatment), data = na.omit(hirs))
summary(gam4)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ Treatment + s(FGm0, by = Treatment)
##
## Parametric coefficients:
```

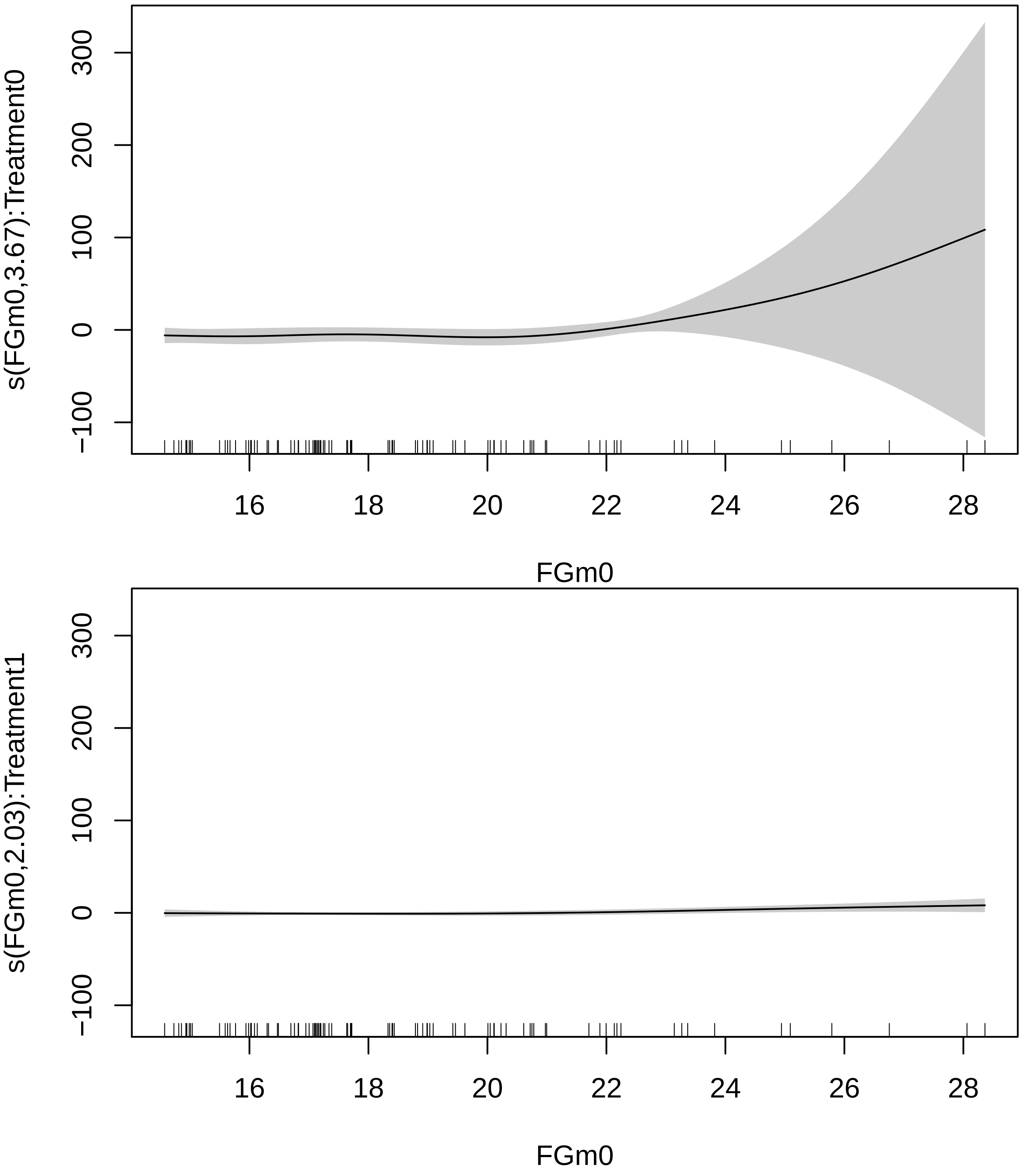
```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.436      4.883   3.571 0.000622 ***
## Treatment1    -9.901      4.975  -1.990 0.050161 .  
## Treatment2    -9.474      4.977  -1.904 0.060763 .  
## Treatment3    -8.389      4.969  -1.688 0.095468 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##                        edf Ref.df     F p-value   
## s(FGm0):Treatment0 3.673  4.376 1.139 0.49681   
## s(FGm0):Treatment1 2.027  2.553 2.659 0.07028 . 
## s(FGm0):Treatment2 1.000  1.000 1.090 0.29984   
## s(FGm0):Treatment3 4.497  5.473 4.198 0.00188 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =  0.291   Deviance explained = 40.3%
## GCV = 23.227  Scale est. = 19.348    n = 91
```

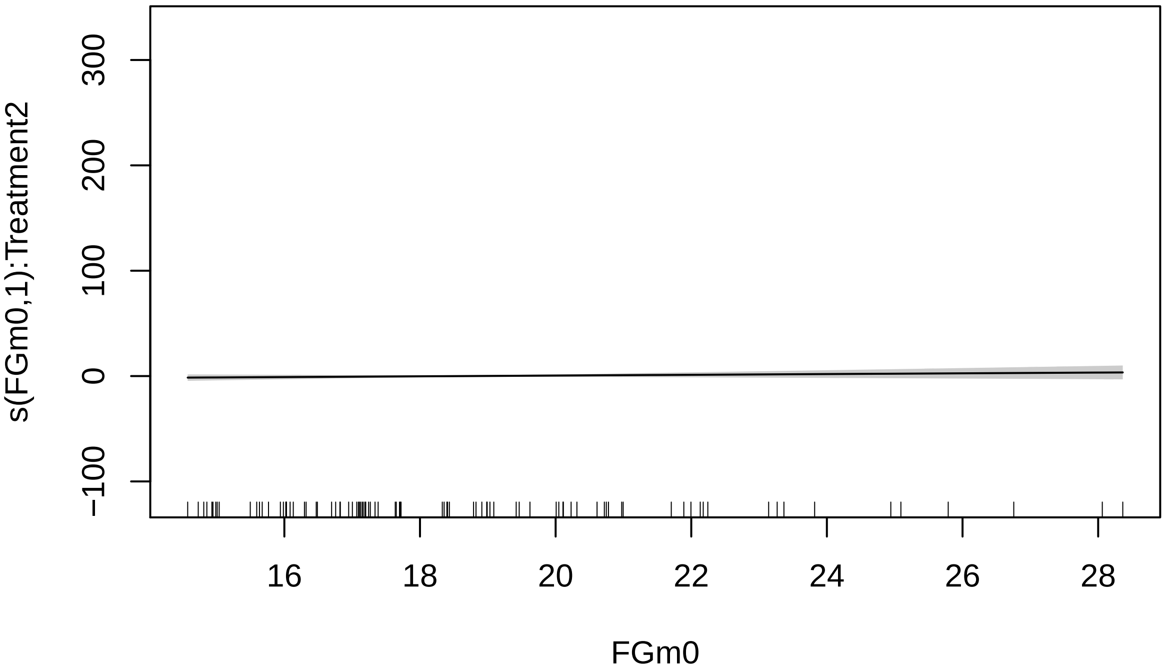
```
anova(gam4, gam2, test = "F")
```

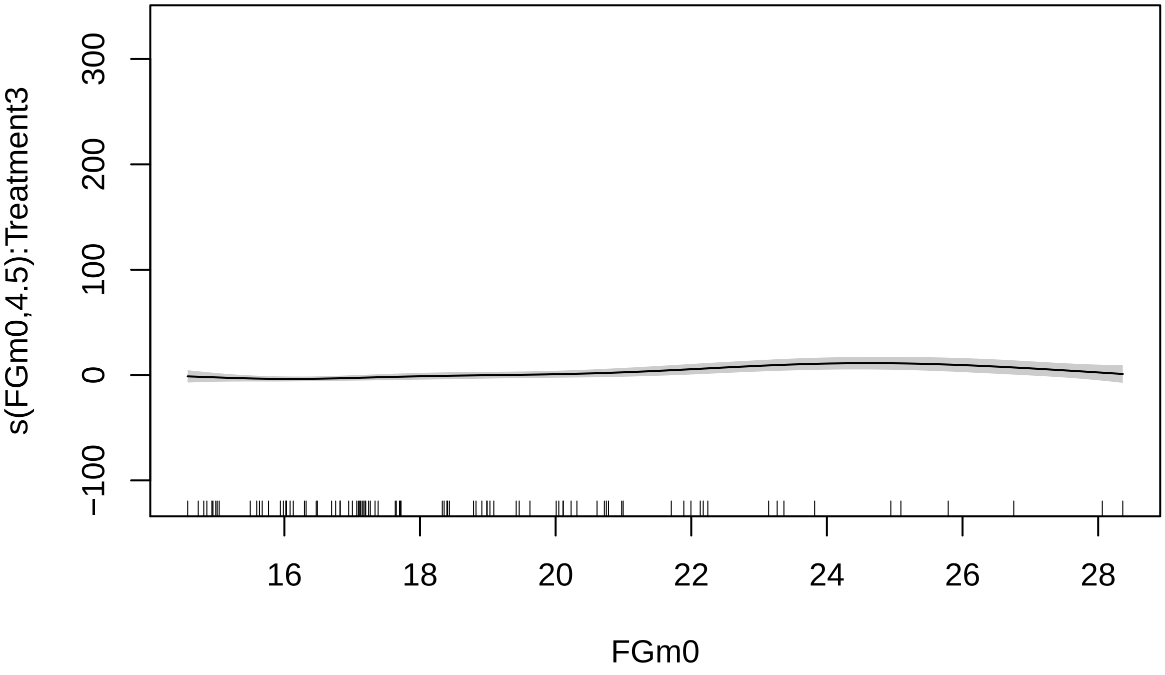
```
## Analysis of Deviance Table
## 
## Model 1: FGm12 ~ Treatment + s(FGm0, by = Treatment)
## Model 2: FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) + 
##     s(DiaPres) + s(weight) + s(height)
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)  
## 1    73.598     1466.6                                 
## 2    68.197     1224.0 5.4012   242.59 2.6031 0.02916 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again we reject the smaller model for the full model. Also we can see that almost all the s(FGm0):Treatment have a significant p value impying that we shoulf fit a model without these.

```
plot(gam4, residuals = TRUE, shade=TRUE, seWithMean=TRUE)
```

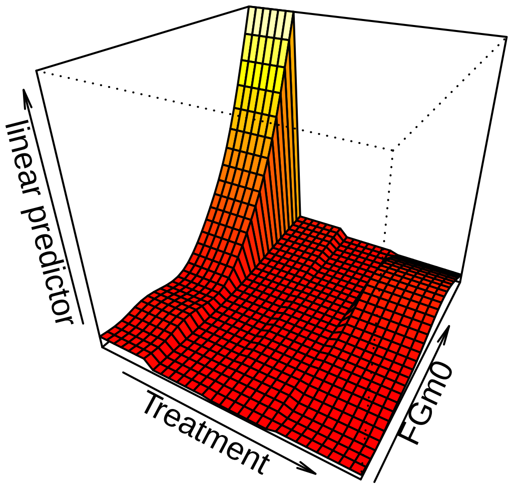

s(FGm0,3.67):Treatment0
−100
0
100
200
300
16
18
20
22
24
26
28
FGm0
s(FGm0,2.03):Treatment1
−100
0
100
200
300
16
18
20
22
24
26
28
FGm0



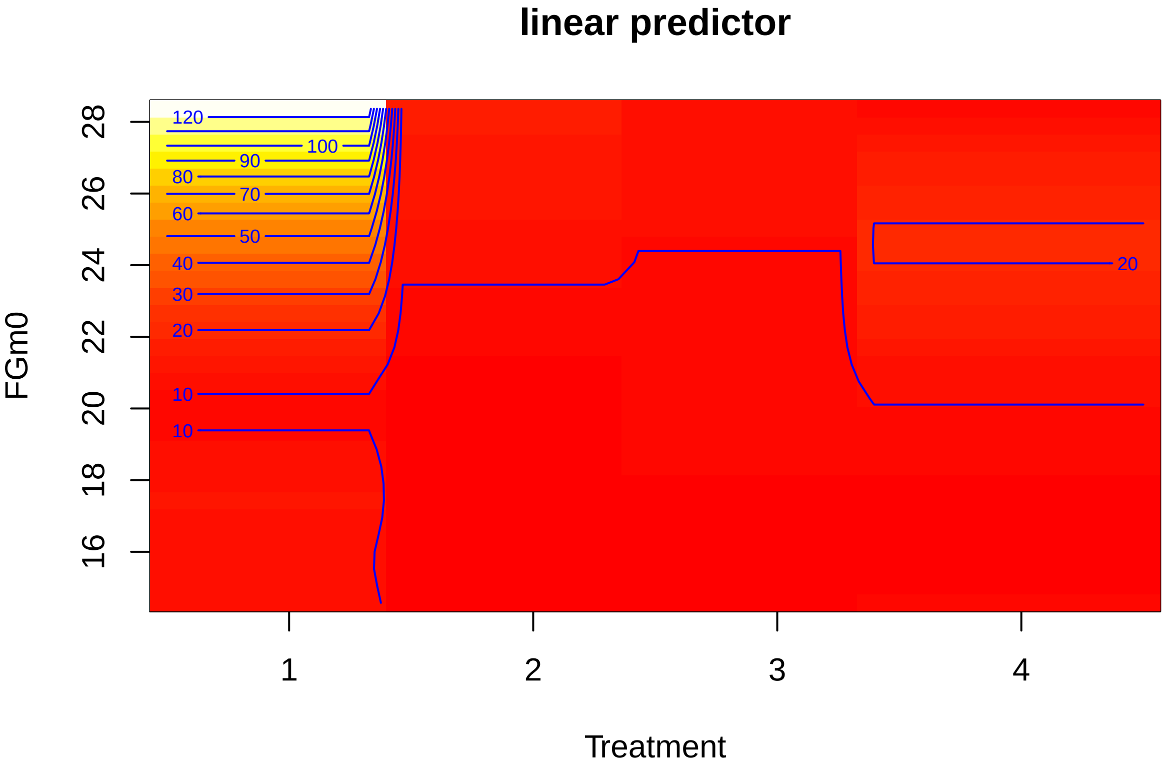


Visualization of the joint effects of variables:

```
vis.gam(gam4, view=c("Treatment","FGm0"), plot.type = "persp", theta=30, phi=30)
```



```
vis.gam(gam4, view=c("Treatment","FGm0"), plot.type = "contour")
```



5th GAM model: smooth model through GAM (FGm12 ~ s(FGm0) + Treatment)

```
gam5 <- gam(FGm12 ~ s(FGm0) + Treatment, data = na.omit(hirs))
summary(gam5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0) + Treatment
```

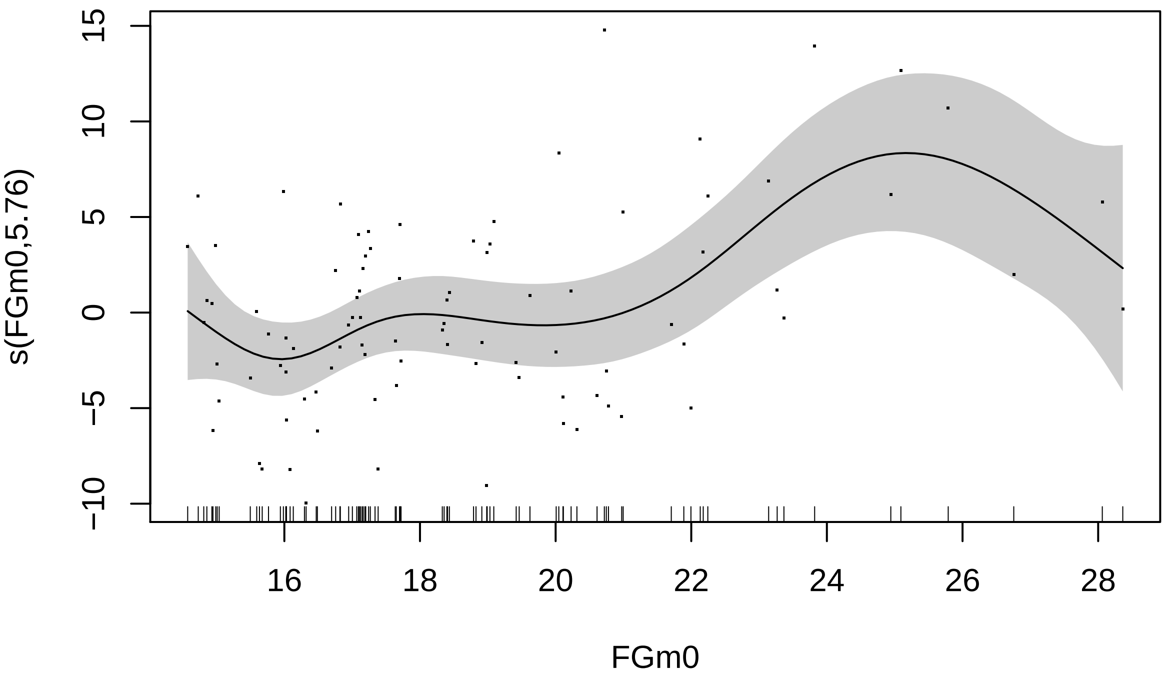
```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.3681     0.9808  12.610  < 2e-16 ***
## Treatment1   -5.0794     1.3986  -3.632 0.000492 ***
## Treatment2   -4.5832     1.3969  -3.281 0.001526 **
## Treatment3   -3.5641     1.3483  -2.643 0.009847 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df     F  p-value
## s(FGm0) 5.763  6.892 3.999 0.000962 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.259   Deviance explained = 33.1%
## GCV = 22.667  Scale est. = 20.235    n = 91
```

```r
anova(gam5, gam2, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0) + Treatment
## Model 2: FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) +
##     s(DiaPres) + s(weight) + s(height)
##   Resid. Df Resid. Dev     Df Deviance     F  Pr(>F)
## 1    80.108     1643.8
## 2    68.197     1224.0 11.911   419.87 2.043 0.03345 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

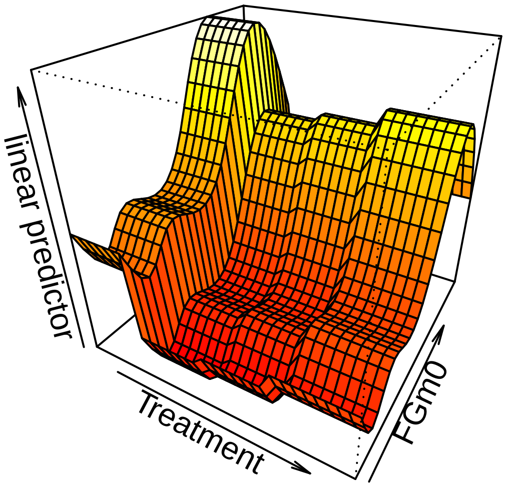
This model is also rejected compared to the full model.

```r
plot(gam5, residuals = TRUE, shade=TRUE, seWithMean=TRUE)
```

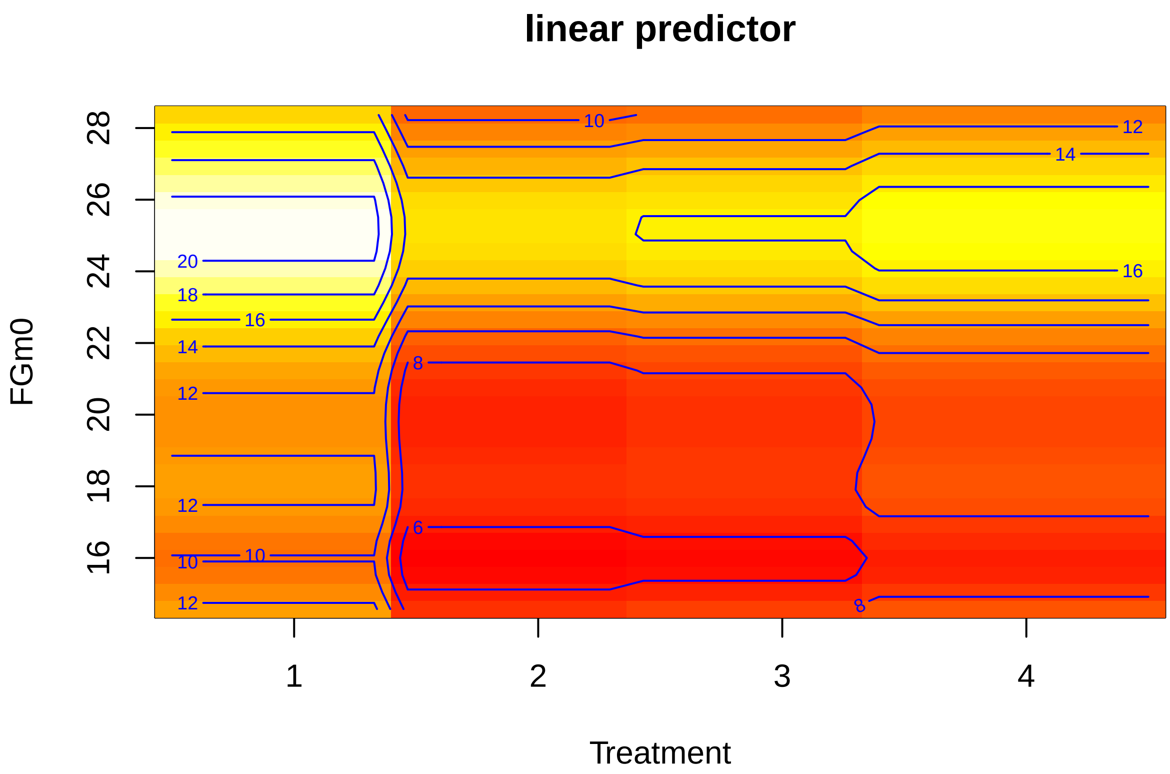


Visualization of the joint effects of variables:

```
vis.gam(gam5, view=c("Treatment","FGm0"), plot.type = "persp", theta=30, phi=30)
```



```
vis.gam(gam5, view=c("Treatment","FGm0"), plot.type = "contour")
```



6th GAM model: smooth model through GAM (FGm12 ~ s(FGm0, by = Treatment) + SysPres + Treatment+ s(FGm0, SysPres)

```r
gam6 <- gam(FGm12 ~ s(FGm0, by = Treatment) + SysPres + Treatment+ s(FGm0, SysPres), data = hirs)
summary(gam6)
```

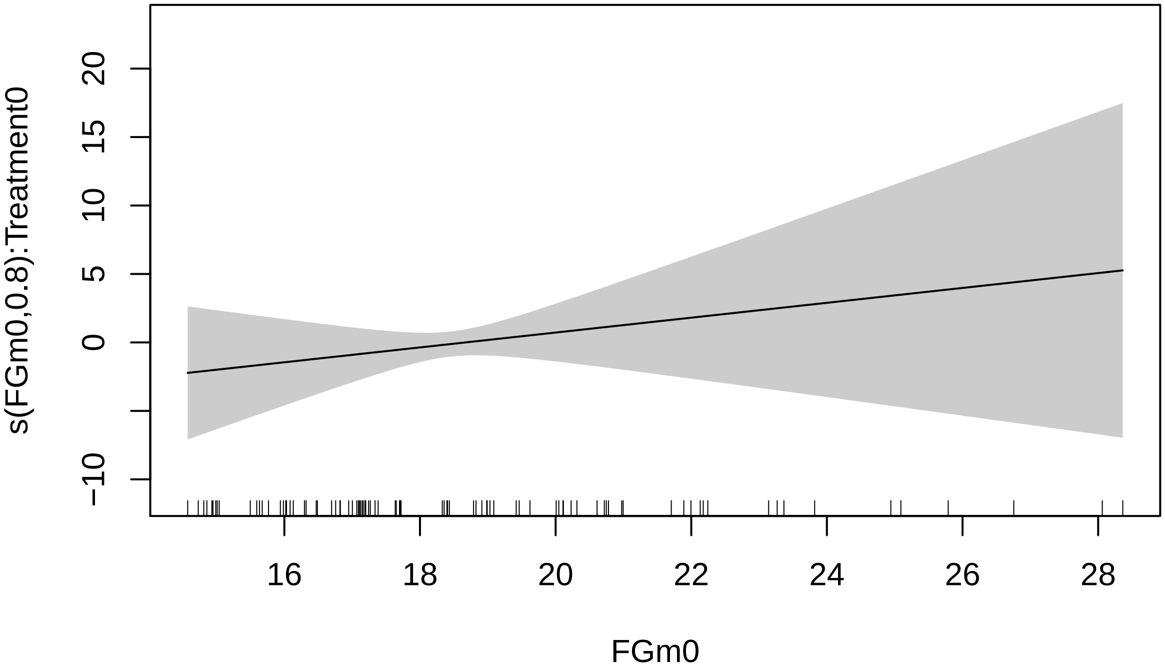
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + SysPres + Treatment + s(FGm0,
##     SysPres)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.286129   0.093700   3.054  0.00313 **
## SysPres      0.096727   0.008976  10.776  < 2e-16 ***
## Treatment1  -4.097486   1.433100  -2.859  0.00551 **
## Treatment2  -3.252065   1.437540  -2.262  0.02660 *
## Treatment3  -2.611464   1.371196  -1.905  0.06070 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                         edf Ref.df     F p-value
## s(FGm0):Treatment0 0.800  0.800 0.974  0.3801
## s(FGm0):Treatment1 3.845  4.695 1.830  0.1283
## s(FGm0):Treatment2 0.800  0.800 2.798  0.1388
## s(FGm0):Treatment3 3.484  4.334 2.446  0.0498 *
## s(FGm0,SysPres)    3.526  4.595 4.561  0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Rank: 68/70
## R-sq.(adj) =  0.303   Deviance explained = 42.2%
## GCV = 23.246  Scale est. = 19.039    n = 91
```

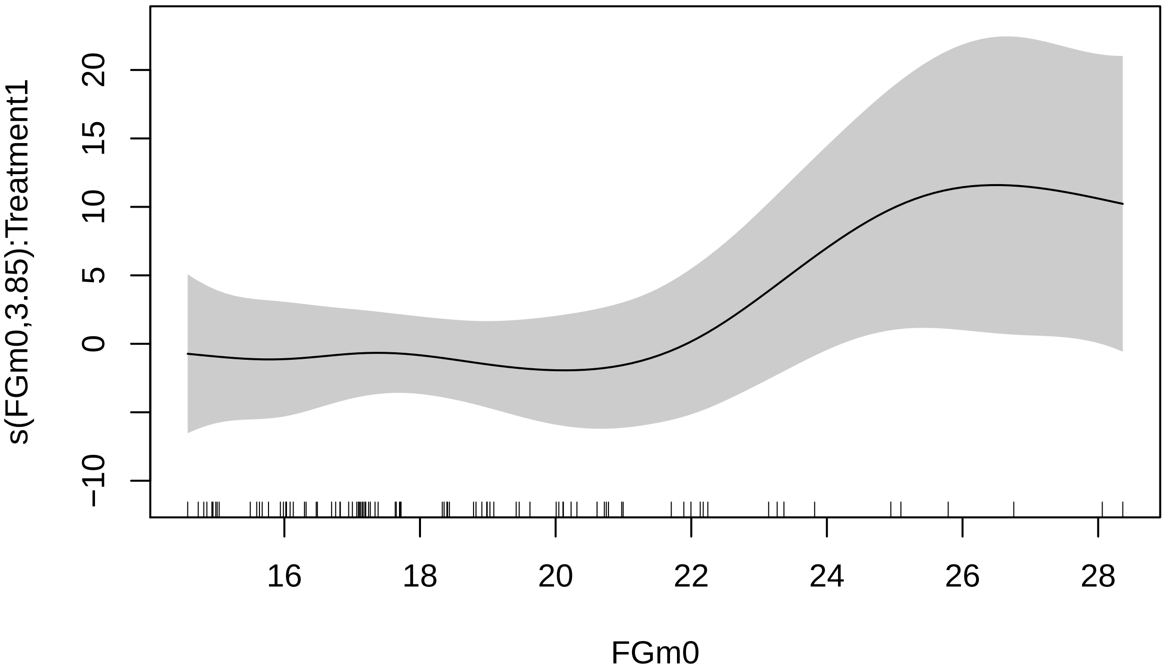
```
anova(gam6, gam2, test = "F")
```

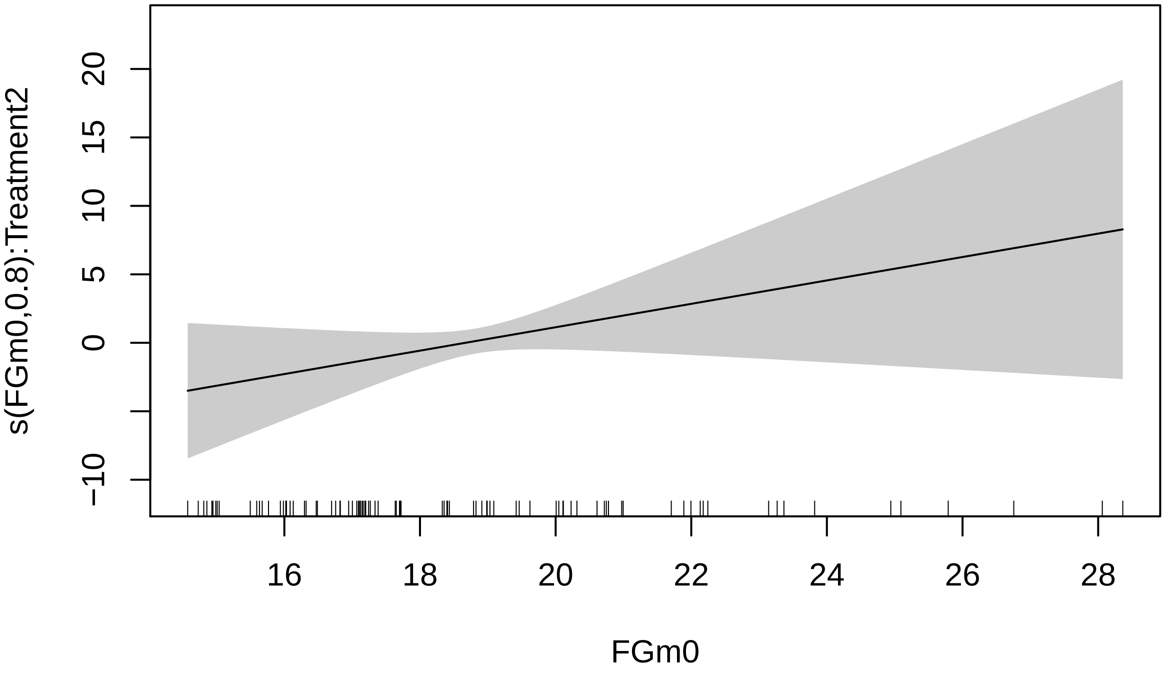
```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0, by = Treatment) + SysPres + Treatment + s(FGm0,
##     SysPres)
## Model 2: FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) +
##     s(DiaPres) + s(weight) + s(height)
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    71.763     1419.1
## 2    68.197     1224.0 3.5659   195.08 3.1706 0.02292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

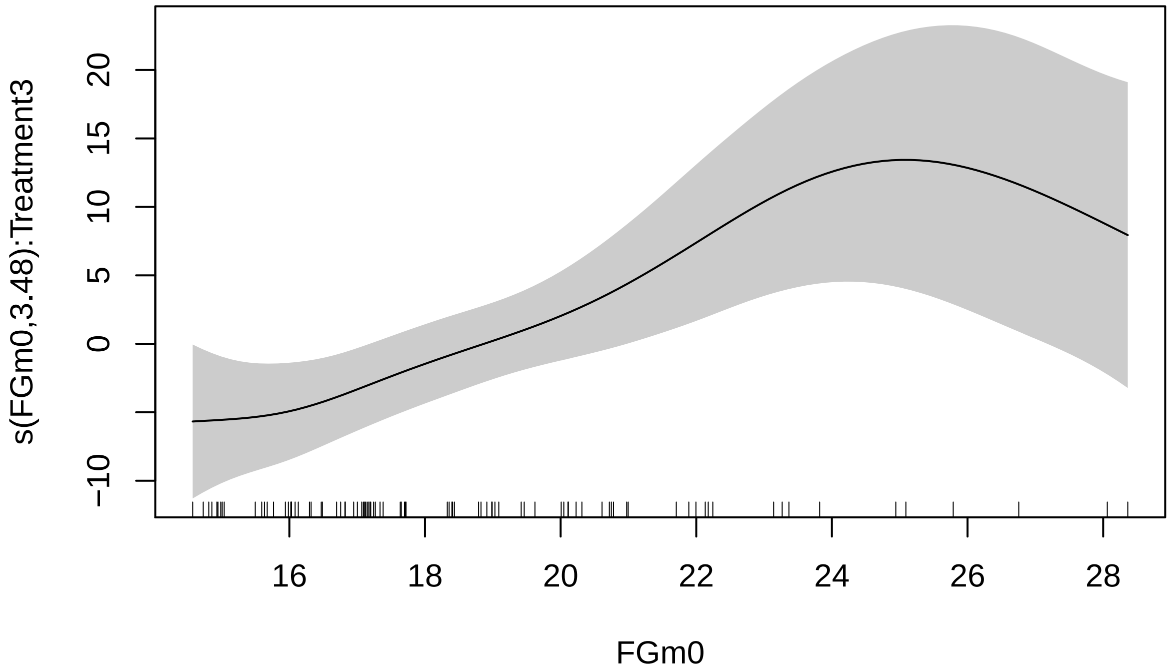
Again the model we prefer is the first full model with all expenatory variables.

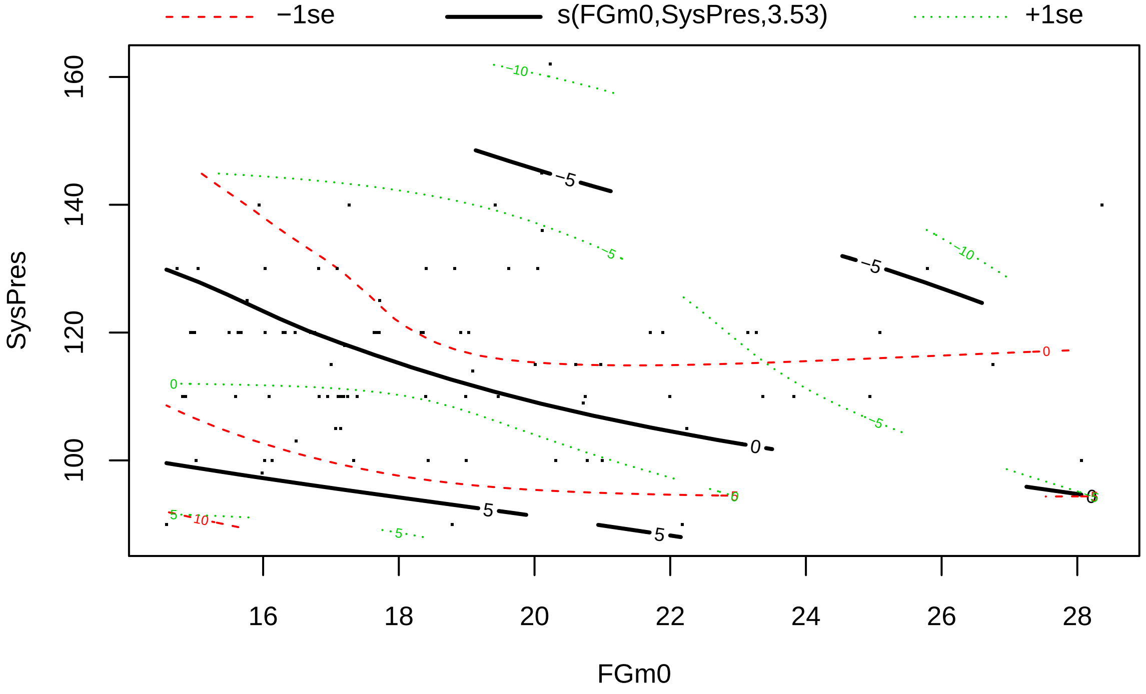
```
plot(gam6, residuals = TRUE, shade=TRUE, seWithMean=TRUE)
```




s(FGm0,3.85):Treatment1
−10
0
5
10
15
20
16
18
20
22
24
26
28
FGm0

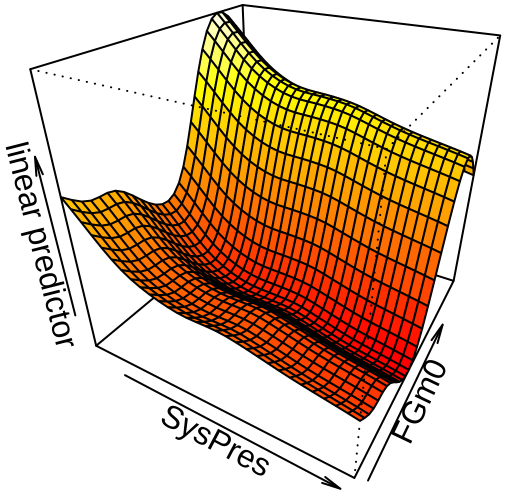

s(FGm0,0.8):Treatment2
−10
0
5
10
15
20
16
18
20
22
24
26
28
FGm0



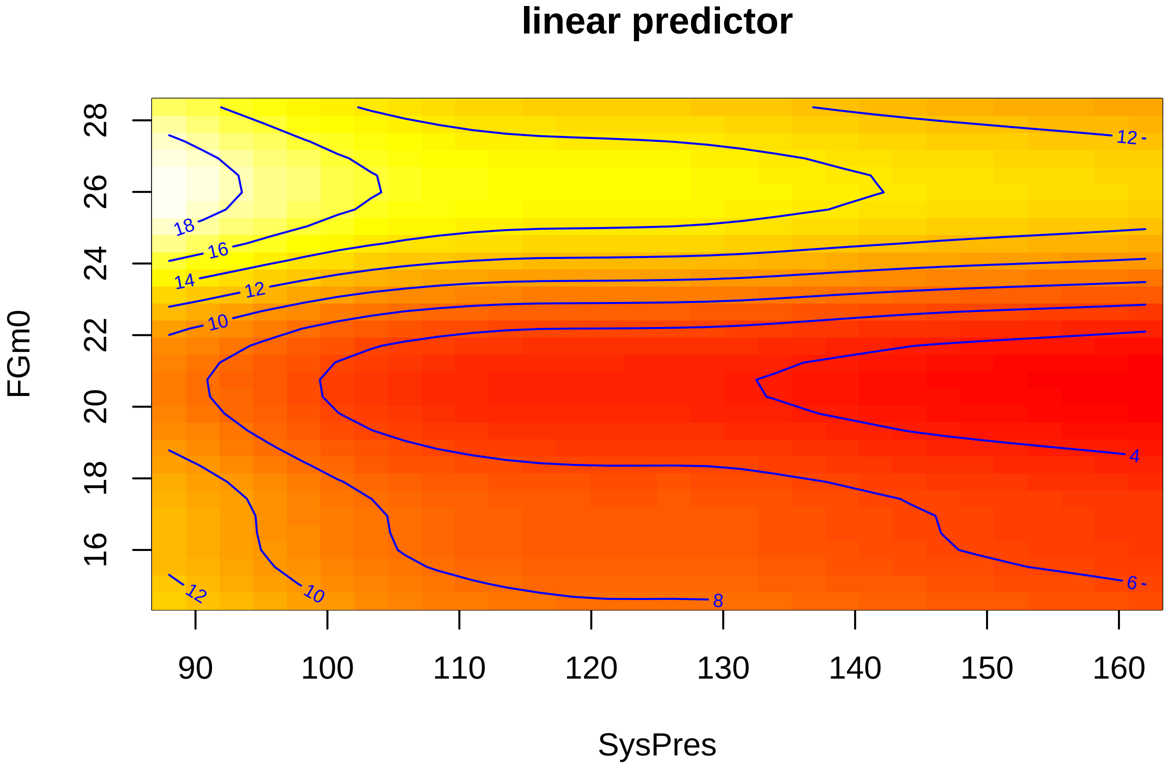


Visualization of the joint effects of variables:

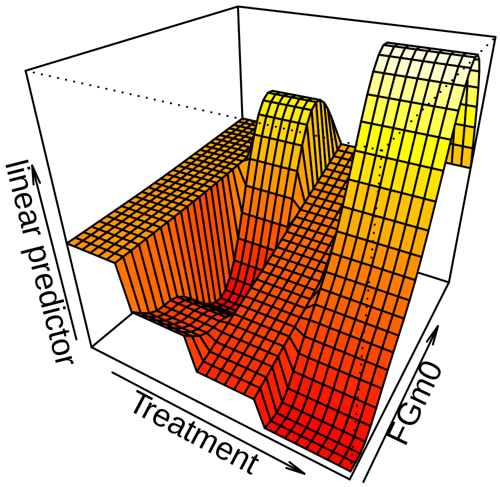
```
vis.gam(gam6, view=c("SysPres","FGm0"), plot.type = "persp", theta=30, phi=30)
```



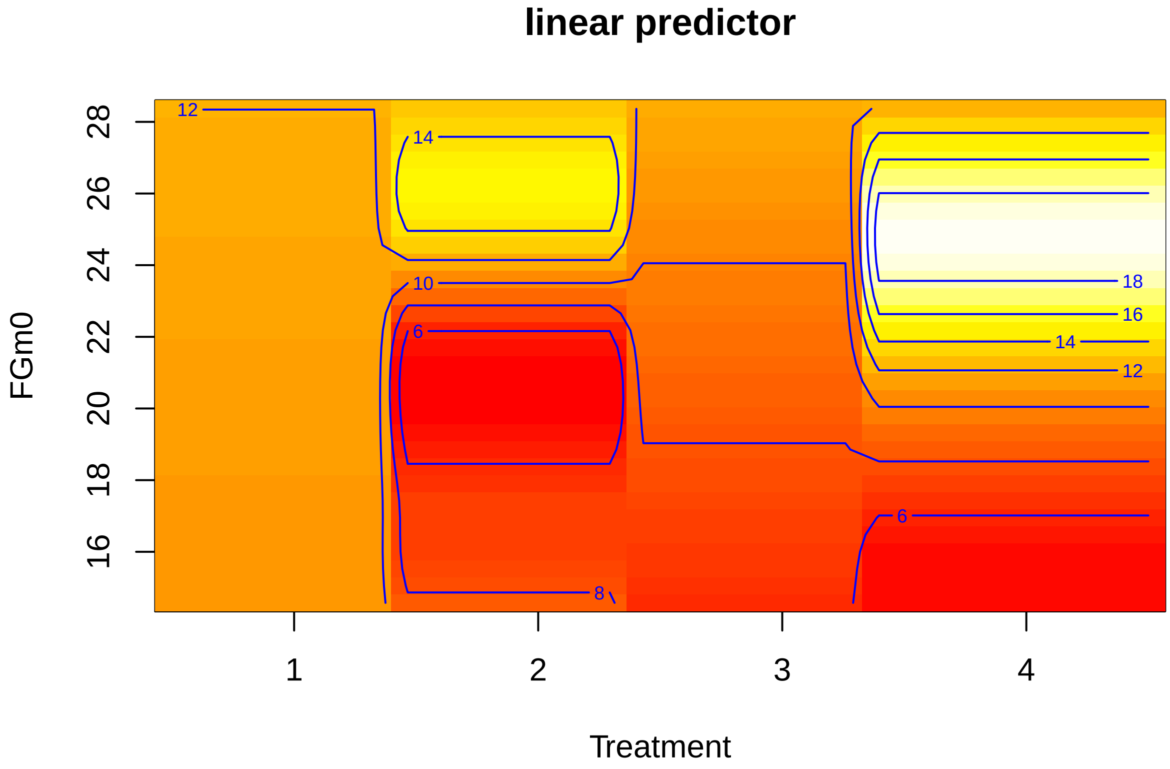
```
vis.gam(gam6, view=c("SysPres","FGm0"), plot.type = "contour")
```



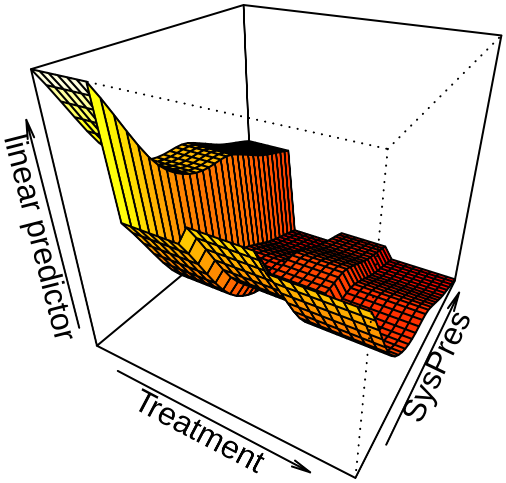
```
vis.gam(gam6, view=c("Treatment","FGm0"), plot.type = "persp", theta=30, phi=30)
```



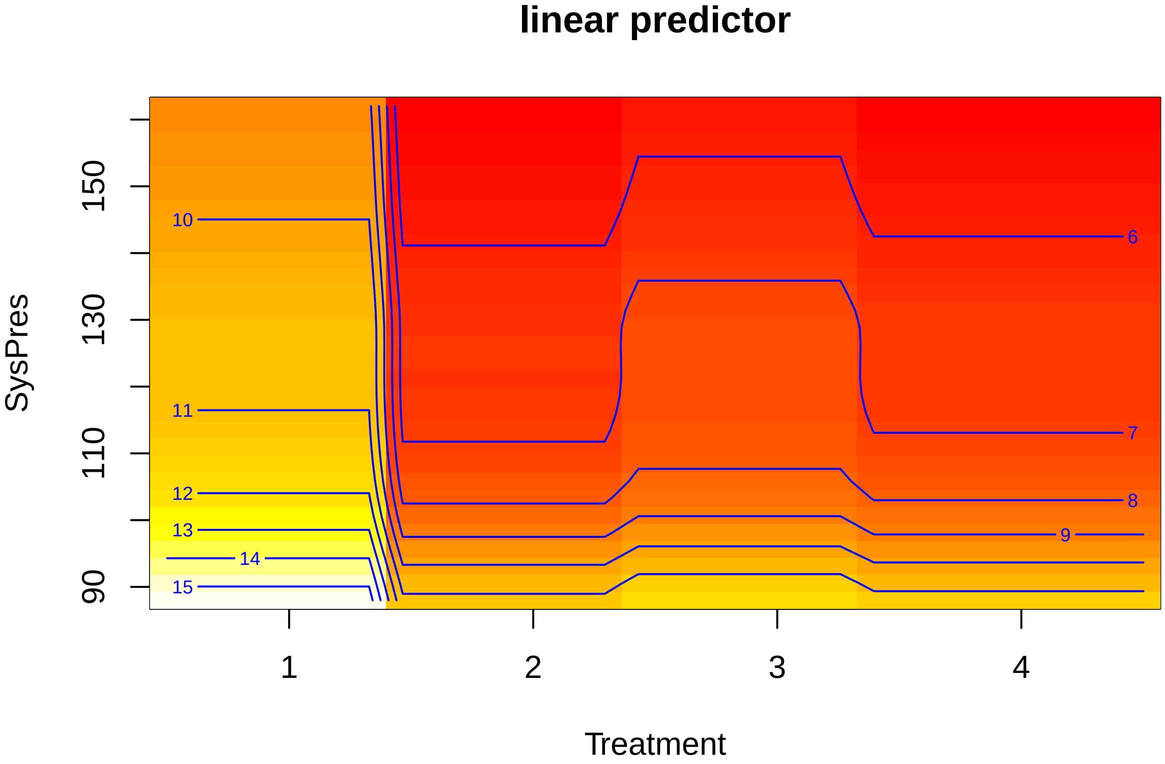
```
vis.gam(gam6, view=c("Treatment","FGm0"), plot.type = "contour")
```



```
vis.gam(gam6, view=c("Treatment","SysPres"), plot.type = "persp", theta=30, phi=30)
```



```r
vis.gam(gam6, view=c("Treatment","SysPres"), plot.type = "contour")
```



ANOVA type tests for the smaller models.

```r
anova(gam1,gam2,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ Treatment + FGm0 + SysPres + DiaPres + weight + height
## Model 2: FGm12 ~ Treatment + s(FGm0) + s(FGm0, by = Treatment) + s(SysPres) +
##     s(DiaPres) + s(weight) + s(height)
##   Resid. Df Resid. Dev     Df Deviance      F   Pr(>F)
## 1    82.000     1857.5
## 2    68.197     1224.0 13.803   633.53 2.6601 0.003905 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is significant so model Gam2 explain better the variance than model 1. We conclude it is better to use a gam model than an ordinary linear model.

```
anova(gam4,gam3,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ Treatment + s(FGm0, by = Treatment)
## Model 2: FGm12 ~ Treatment + s(FGm0, by = Treatment) + SysPres
##   Resid. Df Resid. Dev    Df Deviance      F Pr(>F)
## 1    73.598     1466.6
## 2    71.403     1410.2 2.195   56.332 1.3442 0.2679
```

```
anova(gam5,gam4,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0) + Treatment
## Model 2: FGm12 ~ Treatment + s(FGm0, by = Treatment)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    80.108     1643.8
## 2    73.598     1466.6 6.5095   177.28 1.4076 0.2191
```

```
anova(gam5,gam6,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0) + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + SysPres + Treatment + s(FGm0,
##     SysPres)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    80.108     1643.8
## 2    71.763     1419.1 8.3449   224.79 1.4148 0.2027
```

As we can see the the gam 4 is not rejected comparet to gam3. However gam5 is not rejectec compared to gam4 so we consider this a better model. Gam5 is also not rejected compared to gam6. This measn that the best model which is not the full model is gam5 which is FGm12 ~ s(FGm0) + Treatment. This is a very easy model and as we have seen earlier from the summary it only explains around 33 percent of the deviance.

However it is seen like the full gam model which explains the most variability, also just explains around the 50% of the variablity and has a $R^2_{adj} = 0.368$.