GAM Model

Sergi Carol Laura Cebollero Alex Rodriguez

December 30, 2018

0.1 Introduction

The aim of this lab is to understand the GAM model and get an insight into the fitted model. The lab will consist of two different tasks, the first one will be to create different GAM models and see how well they fit the model. The second one will be to choose the best model according to our criteria.

1 Exercise 1

First of all, we are going to read the hirsutism dataset and explore it:

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

summary(hirs)
```

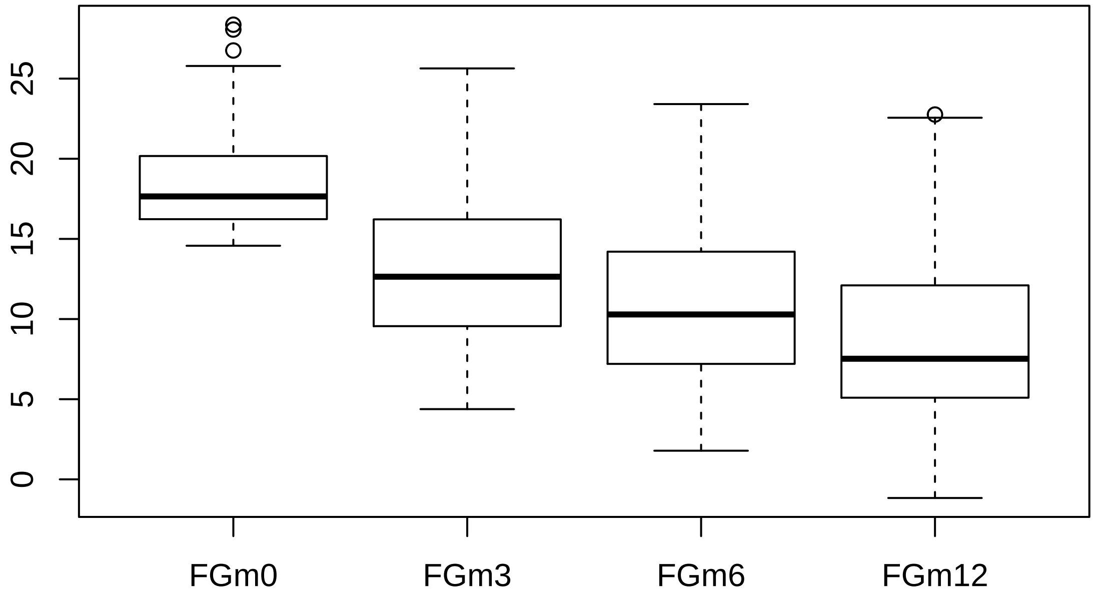
```
##  Treatment      FGm0            FGm3             FGm6       
##  0:23      Min.   :14.57   Min.   : 4.381   Min.   : 1.786  
##  1:26      1st Qu.:16.23   1st Qu.: 9.557   1st Qu.: 7.202  
##  2:24      Median :17.65   Median :12.643   Median :10.286  
##  3:26      Mean   :18.57   Mean   :13.084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

```
head(hirs)
```

```
##   Treatment     FGm0     FGm3      FGm6     FGm12 SysPres DiaPres weight
## 1         0 20.11422 11.16428 12.656592  6.571937     136      71   86.0
## 2         0 16.69666  9.30387  8.530088  9.472867     120      78   52.4
## 3         0 18.78985 14.51187 15.526302 16.109622      90      65   63.0
## 4         0 14.81194 15.67170  9.977178 11.858715     110      70   64.0
## 5         0 17.38289 10.94767  6.161013  4.178238     110      70   62.0
## 6         0 20.97419 10.17097 11.088902  6.920879     115      50   48.5
##   height
## 1   1.71
## 2   1.58
## 3   1.63
## 4   1.57
## 5   1.60
## 6   1.58
```

```
attach(hirs)
```

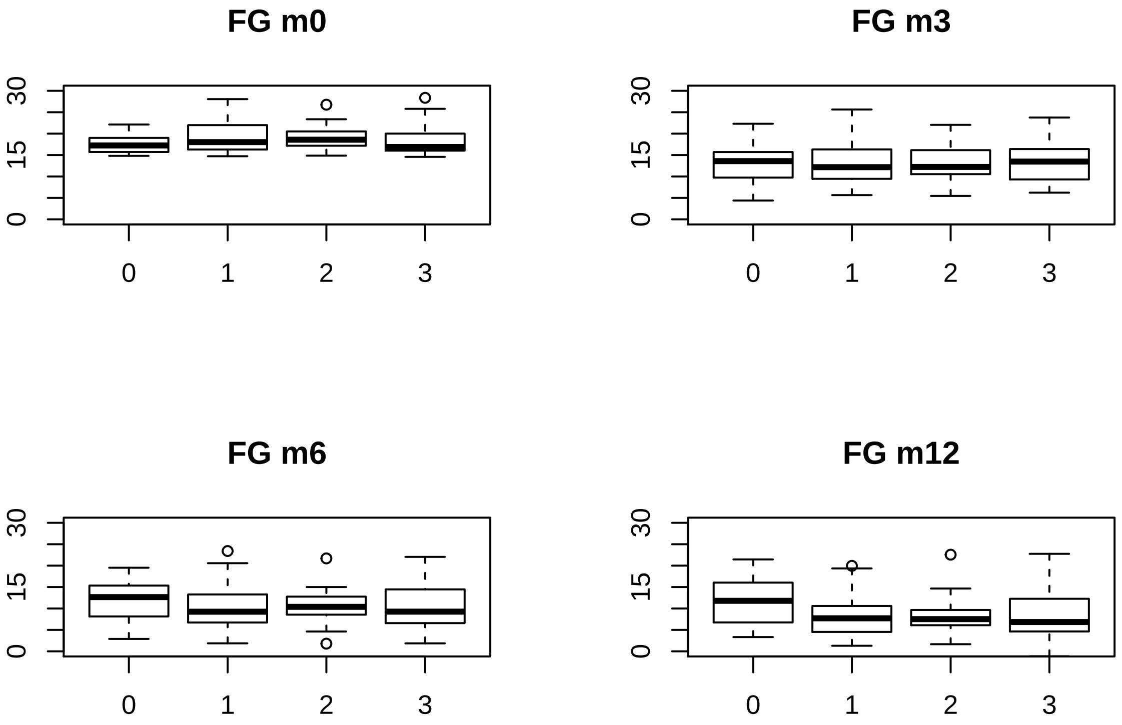
```
boxplot(hirs[,2:5])
```



We can see from the summary and the boxplots that the FG at the start of the treatment (FGm0) is overall greater than afterwards, where the patient has received the treatment.

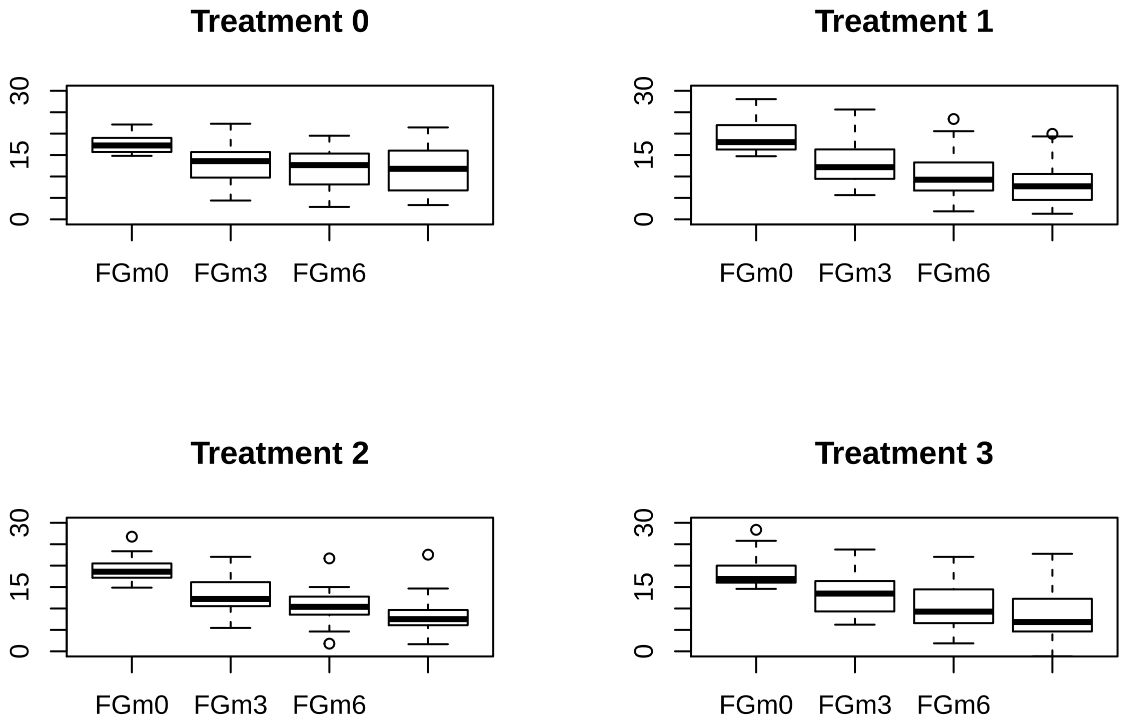
In fact, we can see overall how there is a decrease of the FG as the treatment has an effect on the patient's hirsutism.

```
par(mfrow=c(2,2))
boxplot(hirs[,2]~Treatment,ylim=c(0,30), main="FG m0")
boxplot(hirs[,3]~Treatment,ylim=c(0,30), main="FG m3")
boxplot(hirs[,4]~Treatment,ylim=c(0,30), main="FG m6")
boxplot(hirs[,5]~Treatment,ylim=c(0,30), main="FG m12")
```



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

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


Treatment 0

Treatment 1

Treatment 2

Treatment 3

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

Now if we take a look at the boxplots above where we are comparing the FG at different stages (months 0, 3, 6 and 12) with four different treatments.

We can see how the trend of diminishing hirsutism may vary depending on the treatment the patient receives. For example, with treatment 0 the hirsutism seems to stabilize more than on the other 3 treatments.

On treatment 1, the variance seems way greater than on the other treatments. Treatment 2 seems to not vary a lot and treatment 3 seems to have the most decrease on the hirsutism.

We want to create a model that explains the FGm12 in function of the variables that were mesured at the beginning of the trial.

1.1 Missing values treatment

We can see how there are some NA's in the dataset. Since taking them into account when creating the models will lead to having a diff. number of observations, we are going to remove them. Otherwise, we are not going to be able to perform a comparison of the models created using ANOVA.

```r
nrow(hirs)
```

```
## [1] 99
```

```r
hirs = hirs[complete.cases(hirs), ]
nrow(hirs)
```

```
## [1] 91
```

We can see how 8 rows have been removed. Since there were 8 NAs in weight, height, DiaPres and SysPres, and 8 rows have been removed, it means there are are 8 observations that have NA in all of those 4 variables.

1.1.1 First model

Let's start with a simple model.

```
am1.0 <- gam(FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment, data=hirs)
am1.0
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment
## Total model degrees of freedom 9
##
## GCV score: 25.13893
```

```
summary(am1.0)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.49686   14.85245   1.313 0.192945
## weight       0.02768    0.04425   0.626 0.533308
## height      -8.71024    9.08570  -0.959 0.340540
## DiaPres      0.03525    0.07115   0.495 0.621652
## SysPres     -0.07570    0.05194  -1.458 0.148787
## FGm0         0.59983    0.16862   3.557 0.000626 ***
## 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 **
## ---
## 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
```

We can see how the weight, height, diaPres and SysPres are not relevant according to the p-value, and the relevant are only FGm0 and the treatments. So let's try to achieve a simpler model without the irrelevant variables.

```
am1.1 <- gam(FGm12 ~FGm0 + Treatment, data=hirs)
am1.1
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ FGm0 + Treatment
## Total model degrees of freedom 5
##
## GCV score: 23.72188
```

```r
summary(am1.1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ FGm0 + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5589     3.0695   0.182 0.855956
## FGm0          0.6247     0.1631   3.829 0.000244 ***
## Treatment1   -4.5853     1.4459  -3.171 0.002104 **
## Treatment2   -4.4336     1.4498  -3.058 0.002969 **
## Treatment3   -3.5982     1.3886  -2.591 0.011231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.179   Deviance explained = 21.5%
## GCV = 23.722  Scale est. = 22.418    n = 91
```

This model does indeed look better and R squared is greater than before.

Now let’s try yet another model with a smooth GAM with FGm0 for each treatment:

```r
am1.2 <- gam(FGm12 ~ s(FGm0, by=Treatment) + Treatment, data=hirs)
am1.2
```

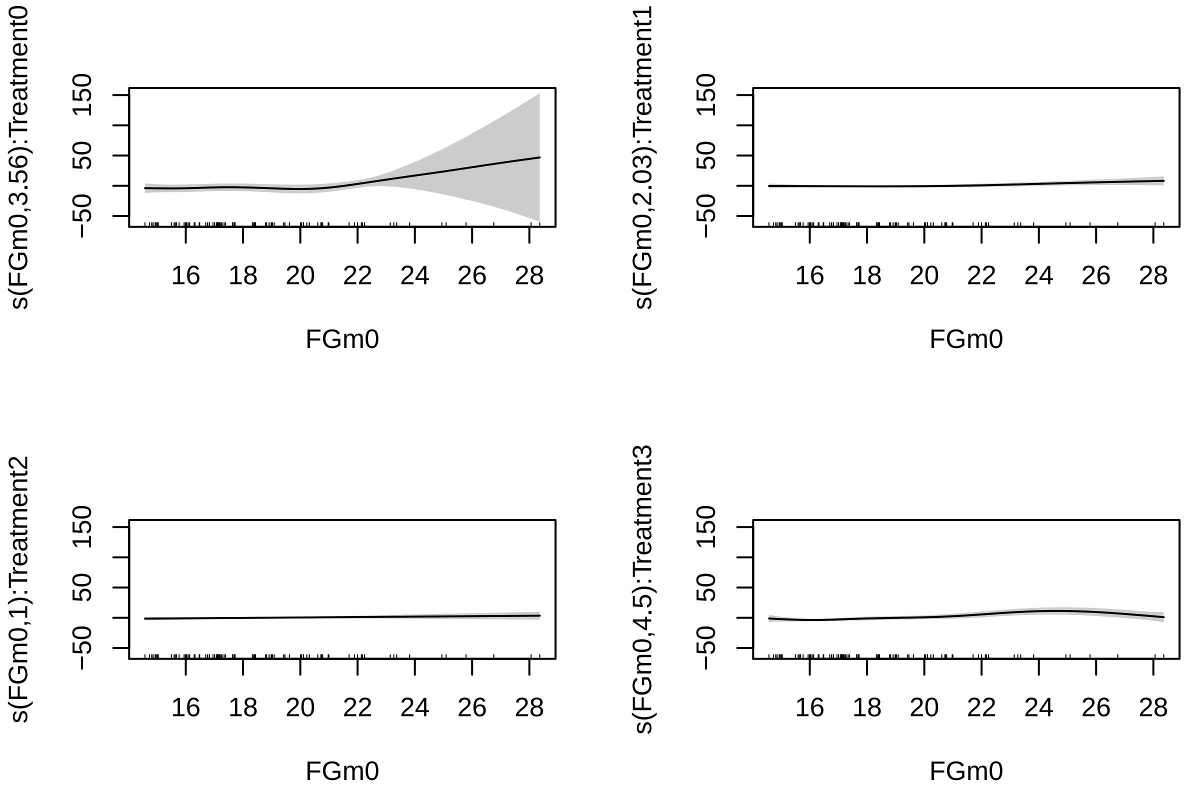
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + Treatment
##
## Estimated degrees of freedom:
## 3.56 2.03 1.00 4.50  total = 15.08
##
## GCV score: 23.24732
```

```r
summary(am1.2)
```

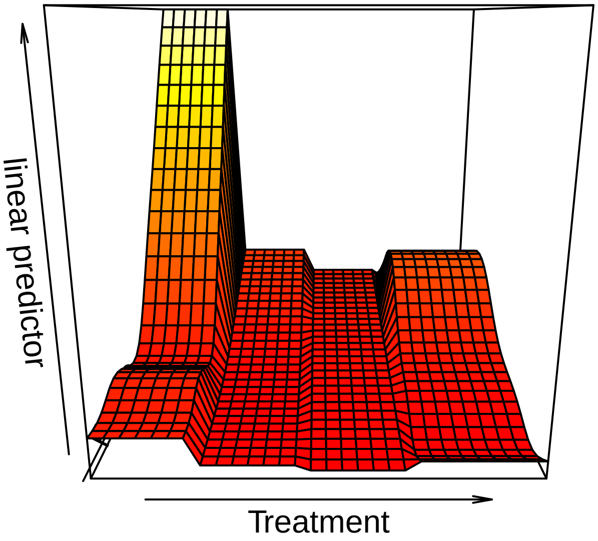
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

```
## FGm12 ~ s(FGm0, by = Treatment) + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   14.965      2.914   5.135 2.12e-06 ***
## Treatment1    -7.430      3.066  -2.423   0.0178 *  
## Treatment2    -7.003      3.070  -2.281   0.0254 *  
## Treatment3    -5.918      3.057  -1.936   0.0566 .  
## ---
## 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 3.557  4.289 0.943 0.57685   
## s(FGm0):Treatment1 2.027  2.553 2.652 0.07073 . 
## s(FGm0):Treatment2 1.000  1.000 1.087 0.30042   
## s(FGm0):Treatment3 4.497  5.473 4.188 0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.29   Deviance explained = 40.1%
## GCV = 23.247  Scale est. = 19.394    n = 91
```

```r
plot.gam(am1.2, page=1, residuals=TRUE, shade=TRUE)
```



```r
vis.gam(am1.2)
```



This model is quite interesting. We can see that Treatment 0 is the only one that does not seem to be linear whereas the other 3 are.

Now we are going to create yet another model with a tensor product smooth for height and weight.

```r
am1.3 <- gam(FGm12 ~ s(FGm0, by=Treatment) + te(weight, height), data=hirs)
am1.3
```

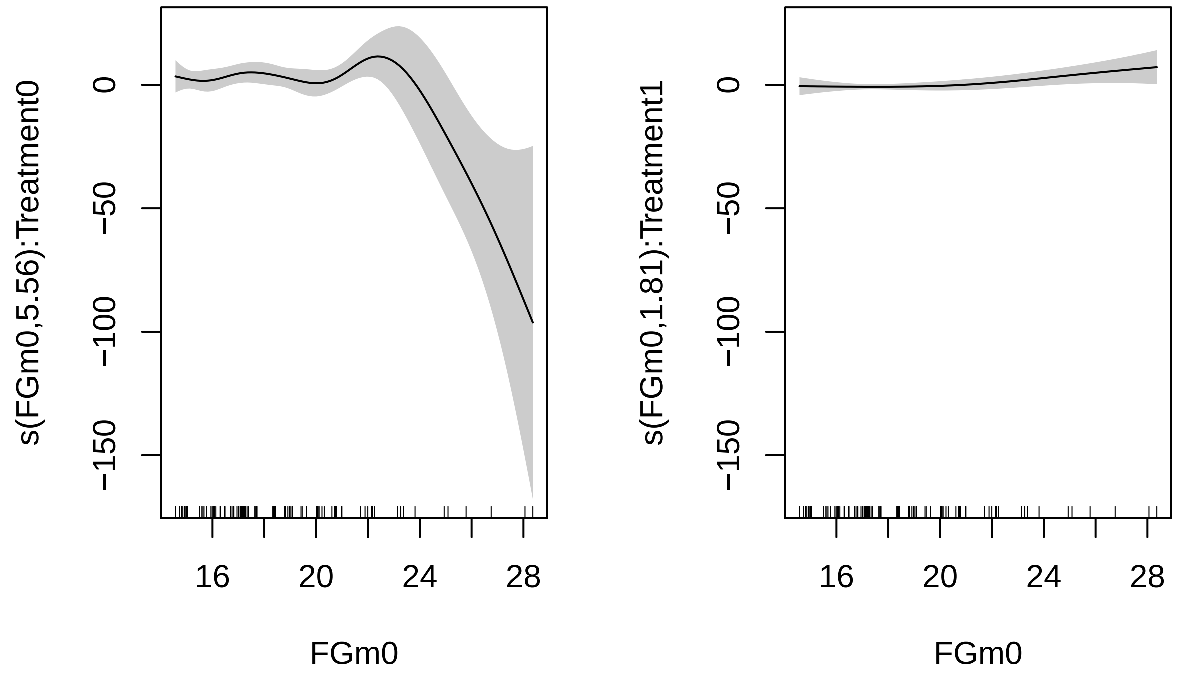
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + te(weight, height)
##
## Estimated degrees of freedom:
## 5.56 1.81 1.00 5.59 3.02  total = 17.99
##
## GCV score: 23.24362
```

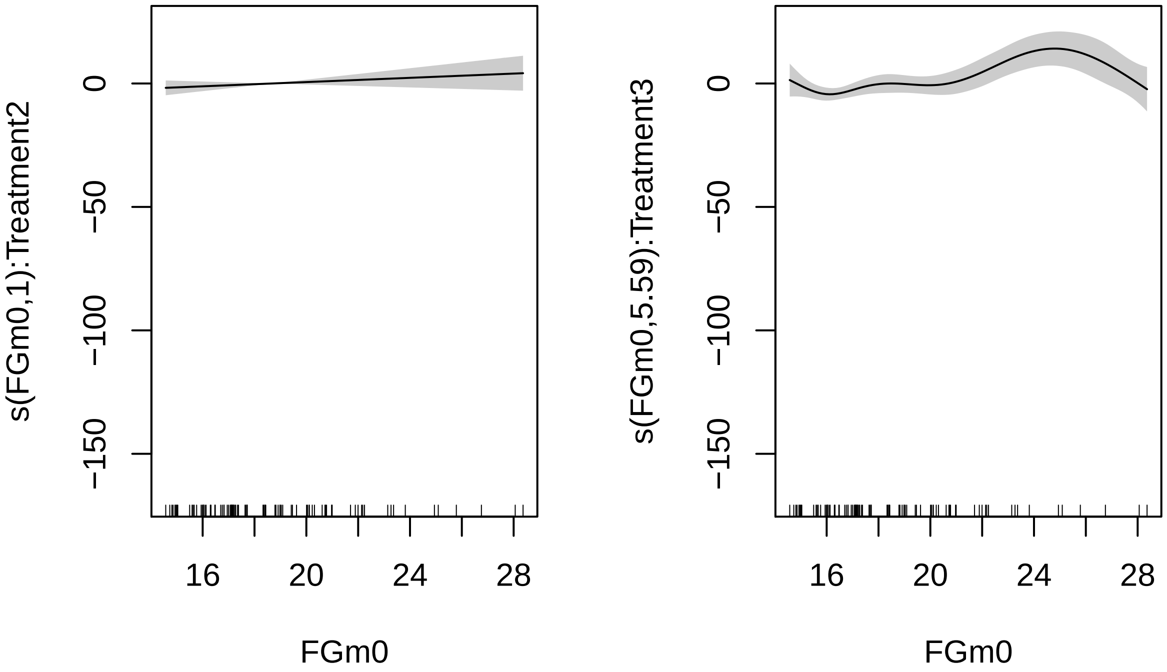
```r
summary(am1.3)
```

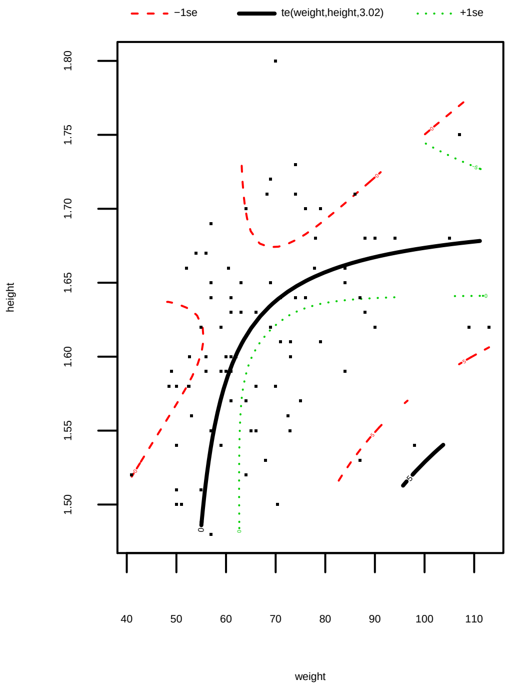
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

```
## FGm12 ~ s(FGm0, by = Treatment) + te(weight, height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2542     0.5386   15.33   <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(FGm0):Treatment0 5.563  6.319 2.687 0.019480 *
## s(FGm0):Treatment1 1.814  2.278 2.422 0.092933 .
## s(FGm0):Treatment2 1.000  1.000 1.386 0.242936
## s(FGm0):Treatment3 5.592  6.654 4.267 0.000685 ***
## te(weight,height)  3.023  3.044 1.309 0.273371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.317   Deviance explained = 44.6%
## GCV = 23.244  Scale est. = 18.648    n = 91
```

```
plot.gam(am1.3, page=3, residuals=TRUE, shade=TRUE)
```




s(FGm0,1):Treatment2
0
−50
−100
−150
16
20
24
28
FGm0
s(FGm0,5.59):Treatment3
0
−50
−100
−150
16
20
24
28
FGm0



Once again it appears that the weight and height are not relevant and we can see that they are not very close in the plot and there are large gaps between the smooths.

Finally, we create the last model, with a smooth for FGm0, weight and height separately.

```r
am1.4 <- gam(FGm12 ~ s(FGm0, by=Treatment) + s(weight) + s(height), data=hirs)
am1.4
```

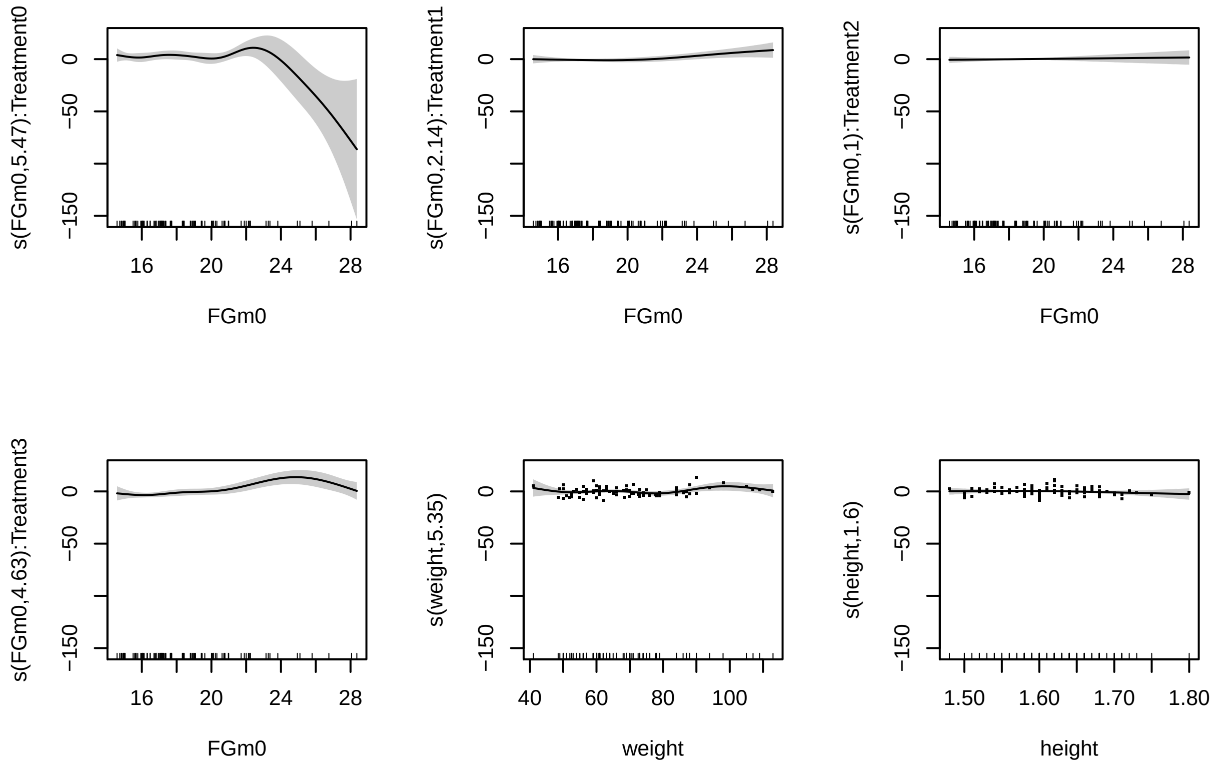
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height)
##
## Estimated degrees of freedom:
## 5.47 2.14 1.00 4.63 5.35 1.60  total = 21.18
##
## GCV score: 22.62655
```

```r
summary(am1.4)
```

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

```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2999     0.5185   16.01   <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(FGm0):Treatment0 5.468  6.235 2.405 0.034806 *
## s(FGm0):Treatment1 2.139  2.676 2.832 0.054001 .
## s(FGm0):Treatment2 1.000  1.000 0.213 0.646211
## s(FGm0):Treatment3 4.627  5.551 4.423 0.000925 ***
## s(weight)          5.348  6.458 1.483 0.188729
## s(height)          1.601  1.968 0.745 0.516870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.364   Deviance explained = 50.7%
## GCV = 22.627  Scale est. = 17.36     n = 91
```

```
plot.gam(am1.4, page=1, residuals=TRUE, shade=TRUE)
```



Yet again, height and weight do not seem that relevant by looking at the p-value. However the R squared adjusted has increased again, which means the proportion of variance explained by this model is greater than before and around almost 37%.

```
am1.5 <- gam(FGm12 ~ s(FGm0, by=Treatment) + s(weight) + s(height) + Treatment, data=hirs)
am1.5
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height) + Treatment
##
## Estimated degrees of freedom:
## 3.37 2.14 1.00 4.81 5.20 1.66  total = 22.18
##
## GCV score: 23.32786
```

```
summary(am1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height) + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.813      2.670   5.548 5.01e-07 ***
## Treatment1    -6.805      2.843  -2.394   0.0194 *
## Treatment2    -6.878      2.841  -2.421   0.0181 *
## Treatment3    -6.124      2.818  -2.173   0.0332 *
## ---
## 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 3.372  4.090 0.812 0.586973
## s(FGm0):Treatment1 2.144  2.681 2.798 0.055861 .
## s(FGm0):Treatment2 1.000  1.000 0.307 0.581204
## s(FGm0):Treatment3 4.808  5.740 4.356 0.000934 ***
## s(weight)          5.201  6.295 1.396 0.224672
## s(height)          1.655  2.042 0.728 0.498997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.354   Deviance explained = 50.6%
## GCV = 23.328  Scale est. = 17.642    n = 91
```

We can see that the deviance explained is now 50.6%, the variance explained is slightly greater and the GCV is OK. However, we can also see that s(weight) and s(height) are not significant according to the p-value, since they are way greater than our confidence interval's p-value, which is 0.05.

2 Task 2

Now we are going to use ANOVA to select the best model that best fits our data at hand and explains the hirsutism at the 12th month.

To do so, we are going to compare each model with the others using ANOVA.

```
anova(am1.0, am1.1,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment
## Model 2: FGm12 ~ FGm0 + Treatment
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1        82     1857.5
## 2        86     1928.0 -4  -70.471 0.7777 0.5428
```

First, we can see that the number of degrees freedom is negative. Thus, we should apply ANOVA reversing the order of the models in the function.

```
anova(am1.1, am1.0,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ FGm0 + Treatment
## Model 2: FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1        86     1928.0
## 2        82     1857.5  4   70.471 0.7777 0.5428
```

We can see how the p-value is way greater than 0.05, so we can establish that the first model am1.1 is better than am1.0.

Let's take a summary of the models:

```
summary(am1.0)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.49686   14.85245   1.313 0.192945
## weight       0.02768    0.04425   0.626 0.533308
## height      -8.71024    9.08570  -0.959 0.340540
## DiaPres      0.03525    0.07115   0.495 0.621652
## SysPres     -0.07570    0.05194  -1.458 0.148787
## FGm0         0.59983    0.16862   3.557 0.000626 ***
## 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 **
## ---
## 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
```

```r
summary(am1.1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ FGm0 + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5589     3.0695   0.182 0.855956
## FGm0          0.6247     0.1631   3.829 0.000244 ***
## Treatment1   -4.5853     1.4459  -3.171 0.002104 **
## Treatment2   -4.4336     1.4498  -3.058 0.002969 **
## Treatment3   -3.5982     1.3886  -2.591 0.011231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.179   Deviance explained = 21.5%
## GCV = 23.722  Scale est. = 22.418    n = 91
```

If we take a look at the summary of the models, although the R squared is slightly greater in am1.0 and although the deviance explained is slightly lower in the am1.1, the GCV is better in the am1.1. Also, the model is simpler in am1.1 since we are not taking taking into account the irrelevant variables that we found on am1.0. Thus, that's probably why ANOVA has chosen am1.0 over am1.1.

Now, having established that we prefer model am1.1 over am1.0, we are going to proceed comparing am1.1 with am1.2.

```r
anova(am1.1,am1.2,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ FGm0 + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + Treatment
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    86.000     1928.0
## 2    73.685     1472.4 12.315   455.59 1.9075 0.04565 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

First of all, the degrees of freedom are positive so we can proceed onto the comparison.

If we stick to an interval confidence of 95%, then the p-value is 0.05, and we can see that the p-value obtained is slightly smaller than 0.05. Which means it is in the limit of accepting that am1.2 model is better.

Let's take a look at the summary of am1.2:

```r
summary(am1.2)
```

```
##
## Family: gaussian
## Link function: identity
```

```
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.965      2.914   5.135 2.12e-06 ***
## Treatment1    -7.430      3.066  -2.423   0.0178 *
## Treatment2    -7.003      3.070  -2.281   0.0254 *
## Treatment3    -5.918      3.057  -1.936   0.0566 .
## ---
## 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 3.557  4.289 0.943 0.57685
## s(FGm0):Treatment1 2.027  2.553 2.652 0.07073 .
## s(FGm0):Treatment2 1.000  1.000 1.087 0.30042
## s(FGm0):Treatment3 4.497  5.473 4.188 0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.29   Deviance explained = 40.1%
## GCV = 23.247  Scale est. = 19.394    n = 91
```

Let's recall that am1.1 had:

```
R-sq.(adj) =  0.179   Deviance explained = 21.5%
GCV = 23.722  Scale est. = 22.418    n = 91
```

We can see that the variance explained, as seen in the R squared adjusted is greater in the am1.2 model. And if we take a look at the deviance explained, we can see that is is twice better explained in am1.2 too. Finally, we can see that the GCV is better also in am1.2. So using the three metrics we can establish that am1.2 model is better than am1.1.

Let's compare am1.2 with am1.3.

```
anova(am1.2,am1.3,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0, by = Treatment) + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + te(weight, height)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    73.685     1472.4
## 2    70.705     1361.5 2.9794   110.93 1.9965 0.1227
```

Again, the degrees of freedom are positive, so we can proceed onto the comparison. The p-value is greater than 0.05, so it's possible that model am1.2 is better than am1.3.

Let's take a look at the summary:

```
summary(am1.3)
```

```
##
## Family: gaussian
## Link function: identity
##
```

```
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + te(weight, height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2542     0.5386   15.33   <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(FGm0):Treatment0 5.563  6.319 2.687 0.019480 *
## s(FGm0):Treatment1 1.814  2.278 2.422 0.092933 .
## s(FGm0):Treatment2 1.000  1.000 1.386 0.242936
## s(FGm0):Treatment3 5.592  6.654 4.267 0.000685 ***
## te(weight,height)  3.023  3.044 1.309 0.273371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.317   Deviance explained = 44.6%
## GCV = 23.244  Scale est. = 18.648    n = 91
```

Remembering the summary of am1.2:

```r
summary(am1.2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.965      2.914   5.135 2.12e-06 ***
## Treatment1    -7.430      3.066  -2.423   0.0178 *
## Treatment2    -7.003      3.070  -2.281   0.0254 *
## Treatment3    -5.918      3.057  -1.936   0.0566 .
## ---
## 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 3.557  4.289 0.943 0.57685
## s(FGm0):Treatment1 2.027  2.553 2.652 0.07073 .
## s(FGm0):Treatment2 1.000  1.000 1.087 0.30042
## s(FGm0):Treatment3 4.497  5.473 4.188 0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.29   Deviance explained = 40.1%
## GCV = 23.247  Scale est. = 19.394    n = 91
```

We can see that the GCV is almost the same, but the R adjusted squared and the deviance explained are slightly better on am1.3. However, a quick glance over the summary of am1.3 tells us that adding te(weight,height) (tensor product smooth) in the model is adding too much complexity on the model for a very small increase on variance and deviance explained. Thus, the tradeoff is too great and ANOVA seems to choose the simple model, which is am1.2.

Thus, we are going to stick with am1.2 as the best model so far, and compare it with am1.4.

```
anova(am1.2,am1.4,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0, by = Treatment) + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height)
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    73.685     1472.4
## 2    66.112     1212.0 7.5727   260.39 1.9808 0.06592 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-value is very close to 0.05 but still greater, so we may consider sticking to am1.2.

Let's take a look at the am1.4 model:

```
summary(am1.4)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.2999     0.5185   16.01   <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(FGm0):Treatment0 5.468  6.235 2.405 0.034806 *
## s(FGm0):Treatment1 2.139  2.676 2.832 0.054001 .
## s(FGm0):Treatment2 1.000  1.000 0.213 0.646211
## s(FGm0):Treatment3 4.627  5.551 4.423 0.000925 ***
## s(weight)          5.348  6.458 1.483 0.188729
## s(height)          1.601  1.968 0.745 0.516870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.364   Deviance explained = 50.7%
## GCV = 22.627  Scale est. = 17.36     n = 91
```

And remembering the summary of am1.2:

```
R-sq.(adj) =   0.29   Deviance explained = 40.1%
GCV = 23.247  Scale est. = 19.394    n = 91
```

We can see that am1.2 has a worse R squared, a worse percentage of deviance explained and a worse GCV. However, they are all slightly worse and, again, it seems that ANOVA sticks to the simpler model, which is am1.2, which takes into account the smooth of FGm0 by Treatment plus the Treatment used, against model am1.4 which actually creates a smooth of weight and height, which adds complexity and seems to be that good of a trade-off. Thus, this can explain ANOVA sticking to am1.2 by a small margin.

Then, we are going to stick also with am1.2 and see how it compares with am1.5, which if we remember, is the same model as am1.4 but with Treatment.

```
anova(am1.2,am1.5,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0, by = Treatment) + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height) + Treatment
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    73.685     1472.4
## 2    65.152     1214.1 8.5324   258.29 1.7159 0.1069
```

Again it seems to stick with am1.2, which means that the Fisher Test tells us to give preference the simplest model, which is am1.2, over am1.5, which is way more complex.

```
summary(am1.5)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + s(weight) + s(height) + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.813      2.670   5.548 5.01e-07 ***
## Treatment1    -6.805      2.843  -2.394   0.0194 *
## Treatment2    -6.878      2.841  -2.421   0.0181 *
## Treatment3    -6.124      2.818  -2.173   0.0332 *
## ---
## 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 3.372  4.090 0.812 0.586973
## s(FGm0):Treatment1 2.144  2.681 2.798 0.055861 .
## s(FGm0):Treatment2 1.000  1.000 0.307 0.581204
## s(FGm0):Treatment3 4.808  5.740 4.356 0.000934 ***
## s(weight)          5.201  6.295 1.396 0.224672
## s(height)          1.655  2.042 0.728 0.498997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.354   Deviance explained = 50.6%
## GCV = 23.328  Scale est. = 17.642    n = 91
```

And remembering the summary of am1.2:

```
R-sq.(adj) =   0.29   Deviance explained = 40.1%
GCV = 23.247  Scale est. = 19.394    n = 91
```

We can see that the worst model has a higher deviance explained and a higher variance explained, but the

GCV is slightly higher, so in this case the am1.2 GCV is better.

However, for us the real explanation is that am1.2 is a simpler model, and using height and weight is not that good. We have also tried to do the comparision with the chi squared test for anova and have achieved the same results as the F test.

2.1 Comparison conclusions

We have found that the model am1.2 is the best one, which is modeled as:

```
Formula:
FGm12 ~ s(FGm0, by = Treatment) + Treatment
```

Although there were a couple of models that contested this model as being the best, since they had a better percentage of deviance explained as well as a greater variance explained (R adj. squared) and a slightly smaller GCV, with Fisher's test we have selected am1.2.

This is due to this model being simpler and thus, we are not overfitting the model to the data we have.