GAM fits for hirsutism data

Joel Cantero Priego & Ricard Meyerhofer Parra

6/1/2020

Introduction

In this assignment, we are going to use the Hirsutism dataset. Hirsutism is the excessive hairiness on women in those parts of the body where terminal hair does not normally occur or is minimal -for example, a beard or chest hair. The dataset hirsutism.dat contains artificial values of measures corresponding to some patients in this study. The variables are the following:

Variable name	Description	Values
Treatment	Values 0, 1, 2 or 3.	Factor with 4 levels
FGm0	Baseline hirsutism level at the randomization moment	Integer
FGm3	FG value at 3 months	Integer
FGm6	FG value at 6 months	Integer
FGm12	FG value at 12 months, the end of the trial	Integer
SysPres	Baseline systolic blood pressure	Integer
DiaPres	Baseline diastolic blood pressure	Integer
weight	Baseline weight	Integer
height	Baseline height	Integer

(Note: The term baseline means that these variables were measured at the beginning of the clinical trial).

The main objective of this project is to fit several GAM models (including semiparametric models) explaining FGm12 as a function of the variables that were measured at the beginning of the clinical trial (including FGm0) and Treatment (treated as factor). Using functions summary, plot and vis.gam to get an insight into the fitted models. Then we will use function anova to select among them the model (or models) that we think is (are) the most appropriate.

Before start, we have to remove NA values from our dataset and converting Treatment column as factors.

```
hirsutism <- read.table(file="hirsutism.dat", sep="\t", header = T)
hirsutism <- hirsutism[complete.cases(hirsutism), ]
hirsutism$Treatment <- as.factor(hirsutism$Treatment)
```

Model 1

Our first model will be a linear model predicting FGm12 with just 2 terms: FGm0 and Treatment.

```
model.1 <- gam(FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment, data=hirsutism)
summary(model.1)
```

```
##
## 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
```

As we can see in the summary, weight, height, diaPres and SysPres are p-value>0.05, so the relevant variables are just FGm0 and Treatment1, Treatment2 and Treatment3. So our next model would be this one except irrelevant variables.

Model 2

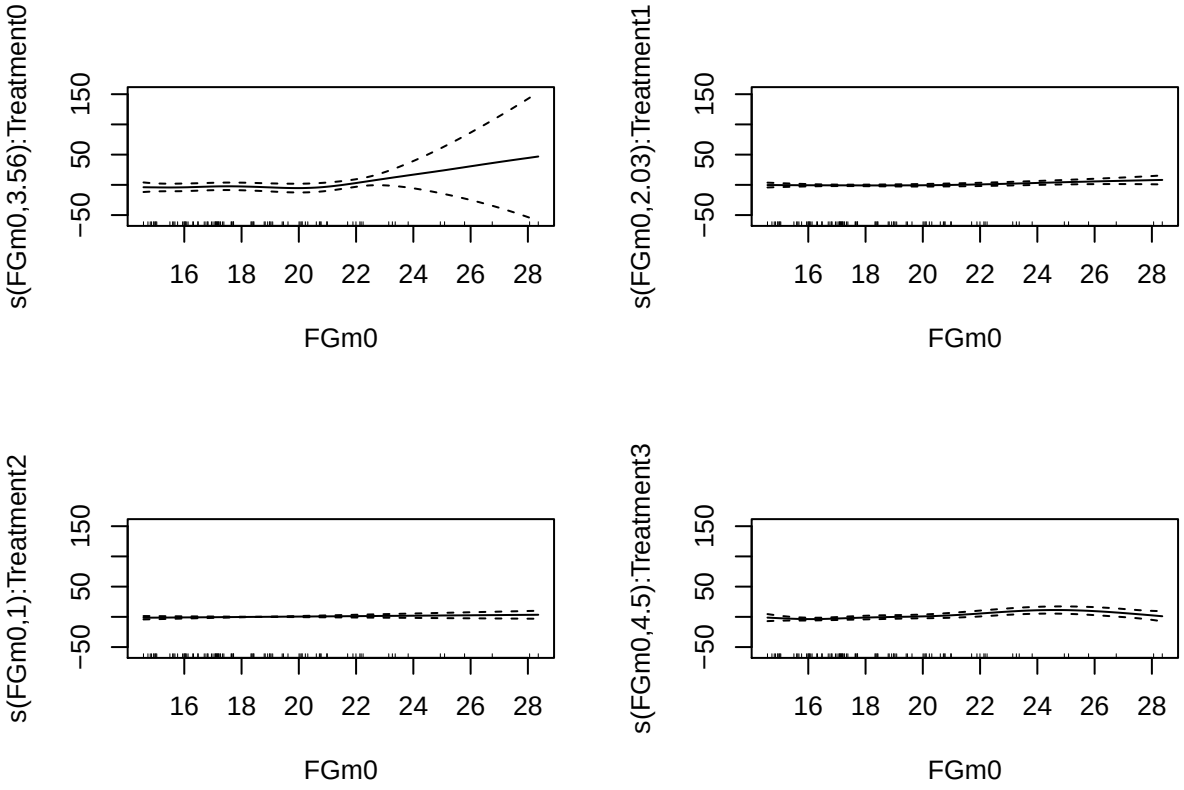
Our second model GAM Model we are going to fit all relevant predictors and we are going to smooth Fgm0 for each Treatment factor.

```
model.2 <- gam(FGm12 ~ s(FGm0, by=Treatment) + Treatment, data=hirsutism)
summary(model.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
```

```
plot(model.2,
     pages=1,
     residuals=TRUE)
```



This model is greater than model 1 because R-sq. (adj) is about 0.291.

Model 3

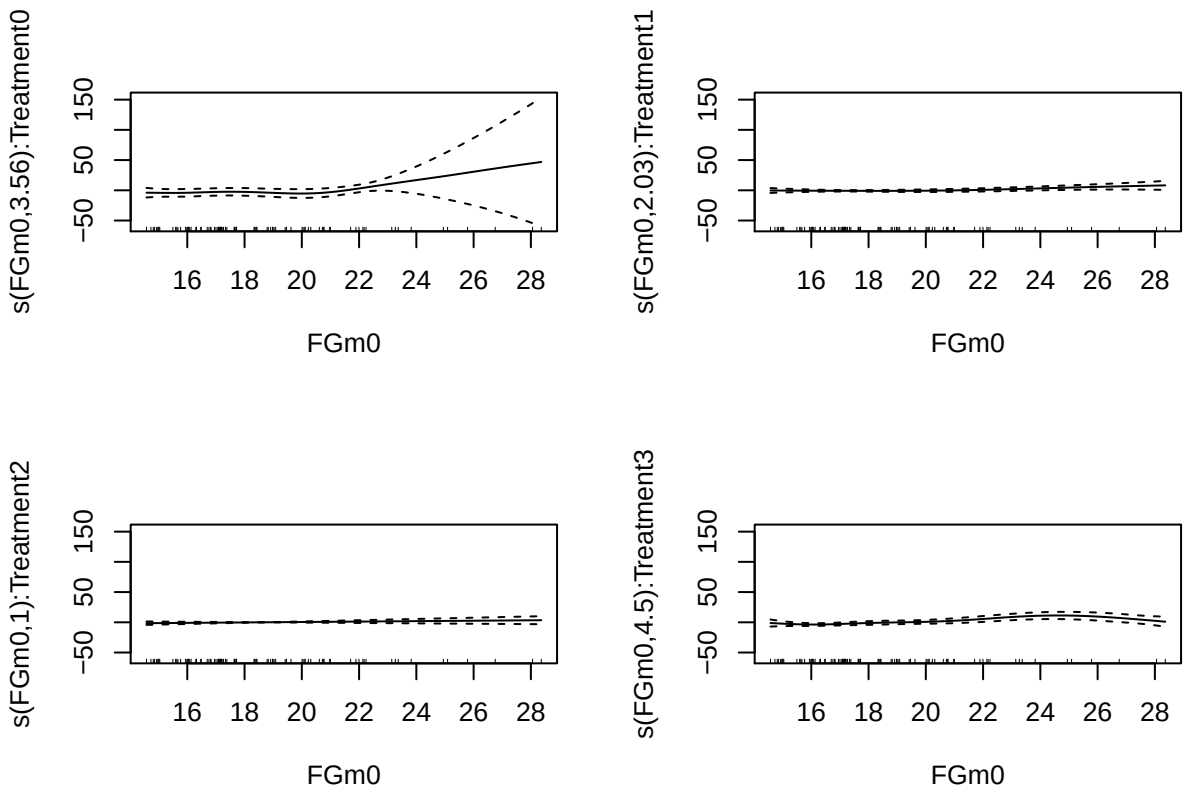
Our next model we are going to smooth Fgm0, weight and height.

```
model.3 <- gam(FGm12 ~ s(FGm0, by=Treatment) + s(weight) + s(height), data=hirsutism)
summary(model.3)
```

```
##
## 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(model.2,
     pages=1,
     residuals=TRUE)
```



We can see again that height and weight are not relevant if we observe p-value. Otherwise, the R squared is about 0.367 and is better than the previous one.

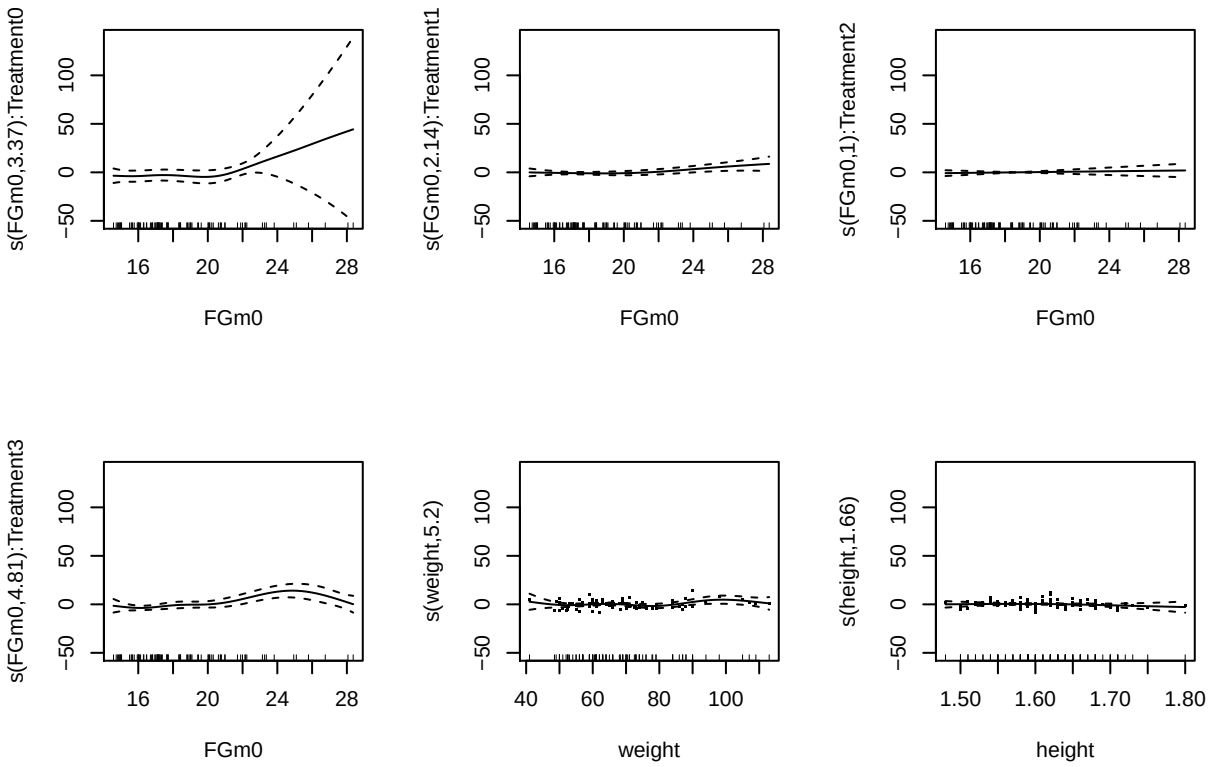
Model 4

Now, we are going to use the previous model adding Treatment

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

```
##
## 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
```

```r
plot(model.4,
     pages=1,
     residuals=TRUE)
```



We can confirm again that weight and height are not relevant and the R-squared is about the previous one (0.354).

Now, having these 4 models, we can comparing them using ANOVA.

Comparing models

Thanks to ANOVA, we are going to select the best comparing each model one by one.

Model 1 vs Model 2

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

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + Treatment
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    82.000     1857.5
## 2    73.685     1472.4 8.3151   385.12 2.3881 0.02249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model.1)
```

```
##
## 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
```

```
summary(model.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 p-value is less than 0.05, so we can confirm that first model is worser than model 2. If we check their R-squared, model 2 has also a greater R-squared.

Model 2 vs Model 3

```
anova(model.2, model.3,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
```

```
summary(model.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
```

```
summary(model.3)
```

```
##
## 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
```

In this case, ANOVA test says that Model 2 is better than model 3 because we get a p-value greater than 0.05 (0.1127).

Model 2 vs Model 4

```
anova(model.2, model.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) + 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
```

```
summary(model.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
```

```
summary(model.4)
```

```
##
## 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
```

Again Fisher Test says that Model 2 is better than model 4 because we get a p-value greater than 0.05 (0.1127).

Conclusion

We can conclude saying that our best model is in the same time one of our simplest models.

FGm12 ~ s(FGm0, by = Treatment) + Treatment

However Model 4 had a better percentage of deviance explained as well as a greater variance explained (R adj. squared), with Fisher's test said that the best one is Model 2.