ASM Practice

GAMs for hirsutism data

Maria Gkotsopoulou & Ricard Monge Calvo & Amalia Vradi

06/01/2020

A clinical trial was conducted to evaluate the effectiveness of an antiandrogen combined with an oral contraceptive in reducing hirsutism for 12 consecutive months. The data set `hirsutism.dat` contains artificial values of measures corresponding to some patients in this study. The variables are the following:

- `Treatment`, with values 0, 1, 2 or 3.
- `FGm0`, it indicates the baseline hirsutism level at the randomization moment (the beginning of the clinical trial). Only women with baseline FG values greater than 15 where recruited.
- `FGm3`, FG value at 3 months.
- `FGm6`, FG value at 6 months.
- `FGm12`, FG value at 12 months, the end of the trial.
- `SysPres`, baseline systolic blood pressure.
- `DiaPres`, baseline diastolic blood pressure.
- `weight`, baseline weight.
- `height`, baseline height.

Our objective 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).

Before proceeding to the model fitting, we remove the 8 rows which have missing values across all additional variables.

During the following discussion, we use the function `gam` from package `mgcv` to fit our additive models.

Baseline linear model

After fitting a complete linear model and removing the non-significant terms, we get a model with the following formula:

- `FGm12 ~ FGm0 + Treatment`

```
##
## 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
```

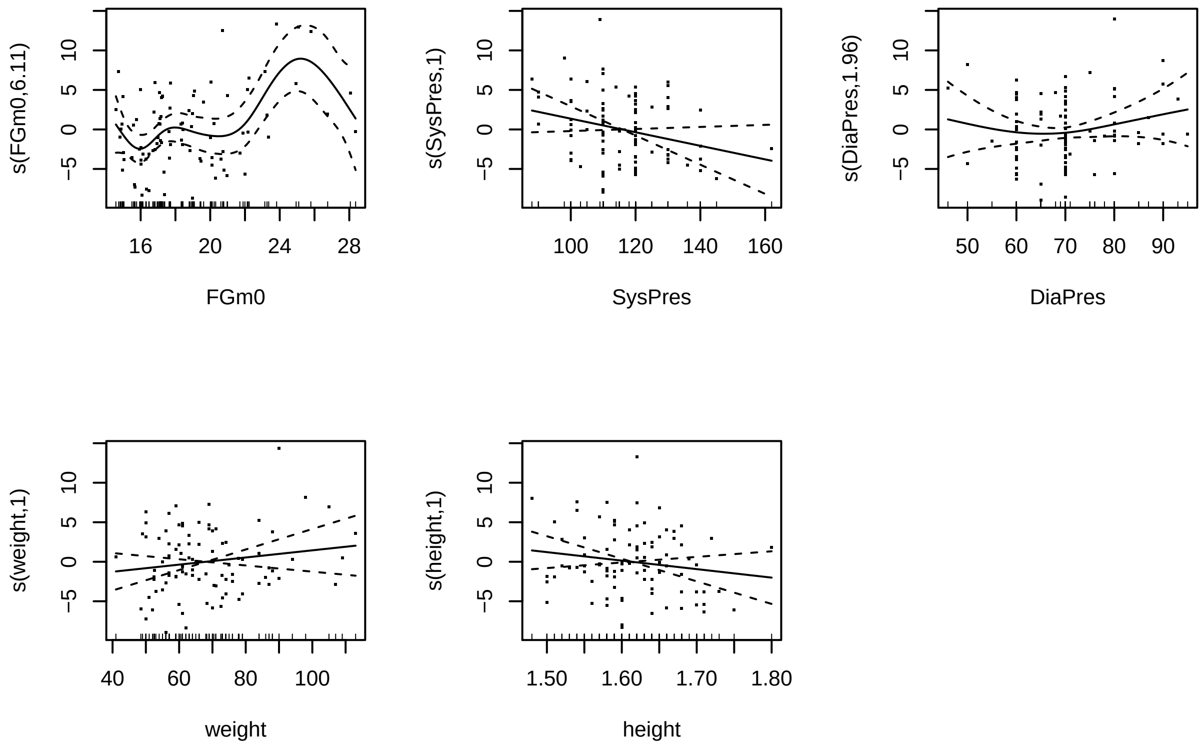
Which we use as a baseline model to compare all following additive models.

Linear vs smooth terms

We want to explore whether our predictors should enter the model as linear parametric or smooth terms. To see this, we fit a model with all numerical predictors as smooth terms, having the following formula:

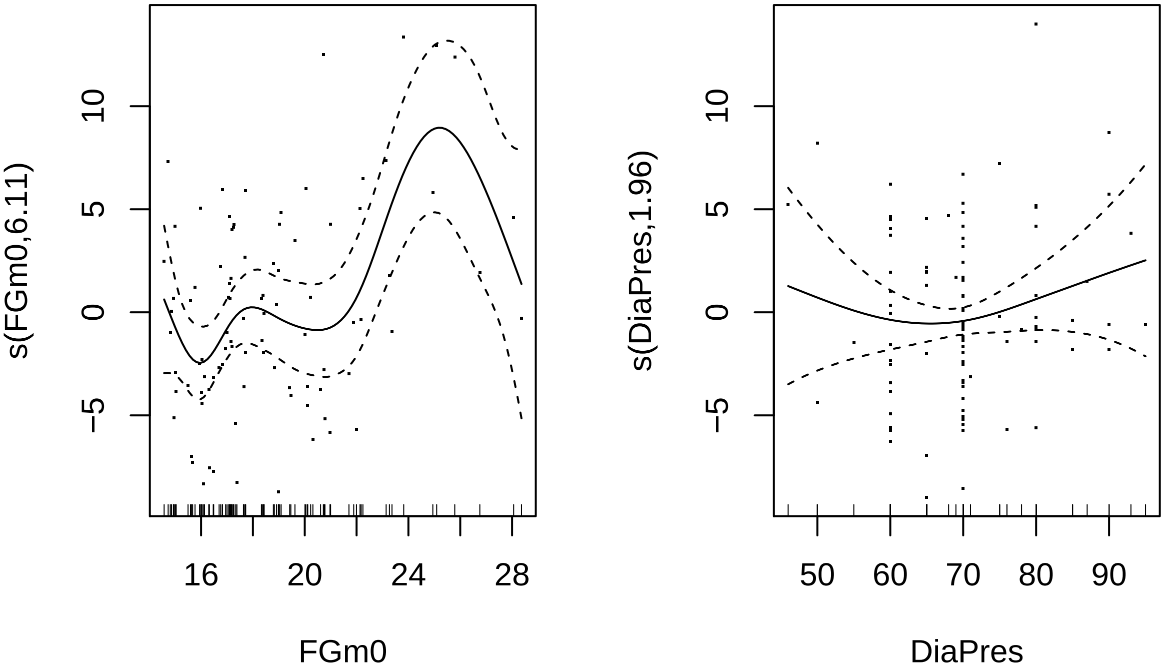
- `FGm12 ~ Treatment + s(FGm0) + s(SysPres) + s(DiaPres) + s(weight) + s(height)`

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ Treatment + s(FGm0) + s(SysPres) + s(DiaPres) + s(weight) +
##     s(height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.477      0.985  12.667  < 2e-16 ***
## Treatment1    -4.868      1.405  -3.465 0.000876 ***
## Treatment2    -4.576      1.418  -3.227 0.001847 **
## Treatment3    -4.154      1.383  -3.005 0.003601 **
## ---
## 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)    6.114  7.238 3.798 0.00136 **
## s(SysPres) 1.000  1.000 3.015 0.08647 .
## s(DiaPres) 1.956  2.464 1.191 0.39964
## s(weight)  1.000  1.000 1.146 0.28767
## s(height)  1.000  1.000 1.454 0.23156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.288   Deviance explained = 39.9%
## GCV =  23.31  Scale est. = 19.45     n = 91
```



From the summary output we see that the effective degrees of freedom (edf) of the *SysPres*, *weight* and *height* predictors are close to 1. In addition, from the plots we see that the form of the effect of these variables is linear. Therefore, we include these terms linearly and refit.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0) + Treatment + SysPres + s(DiaPres) + weight +
##     height
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.68874   14.90229   2.462 0.016086 *
## Treatment1   -4.86769    1.40492  -3.465 0.000876 ***
## Treatment2   -4.57580    1.41794  -3.227 0.001847 **
## Treatment3   -4.15419    1.38259  -3.005 0.003601 **
## SysPres      -0.08588    0.04946  -1.736 0.086532 .
## weight        0.04523    0.04224   1.071 0.287701
## height      -10.74851    8.91321  -1.206 0.231598
## ---
## 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)    6.114  7.238 3.798 0.00136 **
## s(DiaPres) 1.956  2.464 1.191 0.39964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.288   Deviance explained = 39.9%
## GCV =  23.31  Scale est. = 19.45     n = 91
```



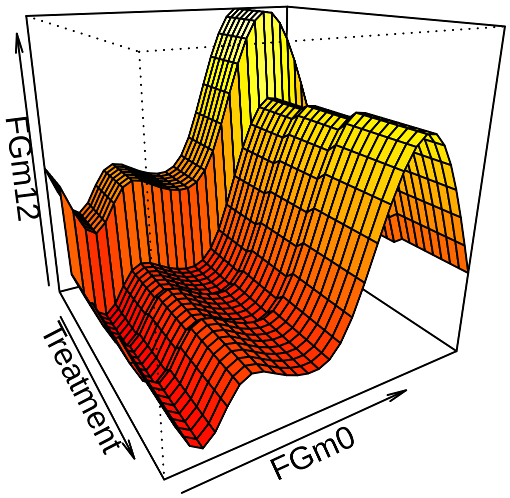
As expected, the model does not change when we turn the smooth terms that had *edf* close to 1 into linear terms.

We now compare this model with the previous linear baseline.

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ FGm0 + Treatment
## Model 2: FGm12 ~ s(FGm0) + Treatment + SysPres + s(DiaPres) + weight +
##     height
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    86.000     1928.0
## 2    74.297     1476.9 11.703   451.12 1.9819 0.0391 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the models do not seem to differ greatly, the explained deviance of the smooth term model, as well as the adjusted R^2 value, is higher. Therefore, we prefer the model with smooth terms.

We see a visualization of the predicted values depending on the predictors *Treatment* and *FGm0*:



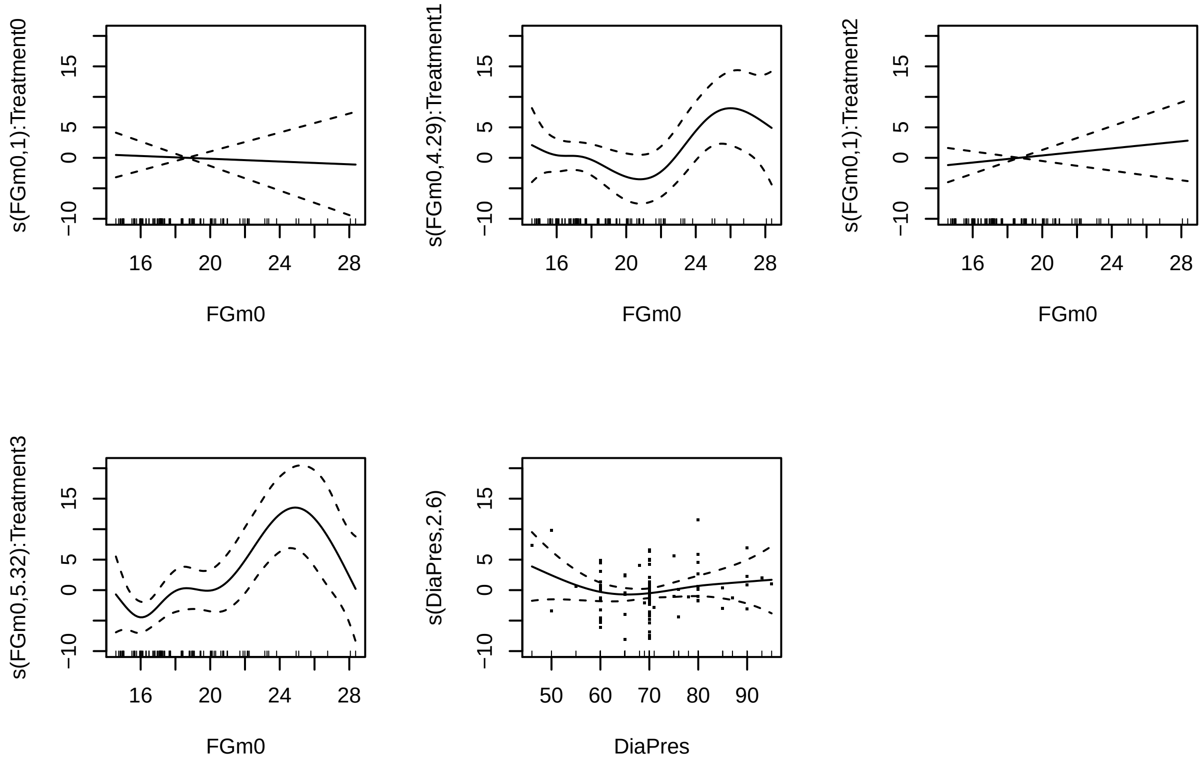
We see that the smooth term *s(FGm0)* has the same shape independently of *Treatment*. We could change the model to get a different estimation for each level of the *Treatment* variable. This could help get more explained deviance and R^2 measures.

Smoothing terms by Treatment

In order to better describe the behaviour of the predictors depending on the *Treatment* value, we can specify the fitting on different smooth terms for each category by using the formula:

- `FGm12 ~ s(FGm0, by= Treatment) + Treatment + SysPres + s(DiaPres) + weight + height`

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + Treatment + SysPres + s(DiaPres) +
##     weight + height
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.20391   14.76871   2.655  0.00983 **
## Treatment1   -3.76443    1.42095  -2.649  0.00997 **
## Treatment2   -3.48534    1.40591  -2.479  0.01559 *
## Treatment3   -2.98611    1.39065  -2.147  0.03525 *
## SysPres      -0.10821    0.05005  -2.162  0.03403 *
## weight        0.07539    0.04223   1.785  0.07855 .
## height      -12.52455    9.03038  -1.387  0.16988
## ---
## 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 1.000  1.000 0.066 0.798678
## s(FGm0):Treatment1 4.294  5.172 2.070 0.075201 .
## s(FGm0):Treatment2 1.000  1.000 0.719 0.399245
## s(FGm0):Treatment3 5.315  6.349 4.531 0.000542 ***
## s(DiaPres)         2.597  3.251 1.199 0.309151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.358   Deviance explained = 50.2%
## GCV = 22.847  Scale est. = 17.523    n = 91
```

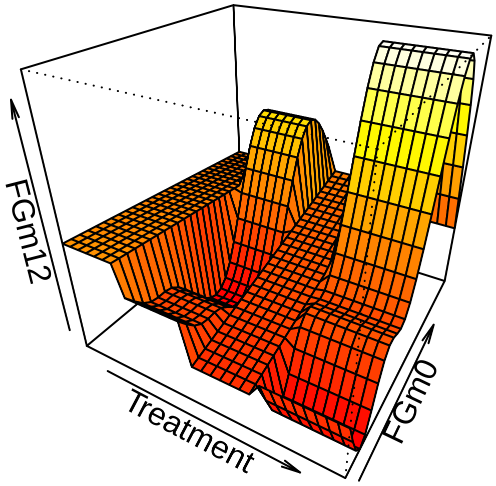


In this case, we see the splines of *s(FGm0)* for *Treatment0* and *Treatment2* are almost linear, while the others are similar as before. We compare with the previous model.

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0) + Treatment + SysPres + s(DiaPres) + weight +
##     height
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + Treatment + SysPres + s(DiaPres) +
##     weight + height
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    74.297     1476.9
## 2    67.228     1223.0 7.0694   253.88 2.0494 0.06077 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even if the models do not seem statistically different in terms of residual deviance, we prefer the latter that has a higher percentage of explained deviance, together with a higher adjusted R^2.

We see a visualization of the predicted values depending on the predictors *Treatment* and *FGm0*:

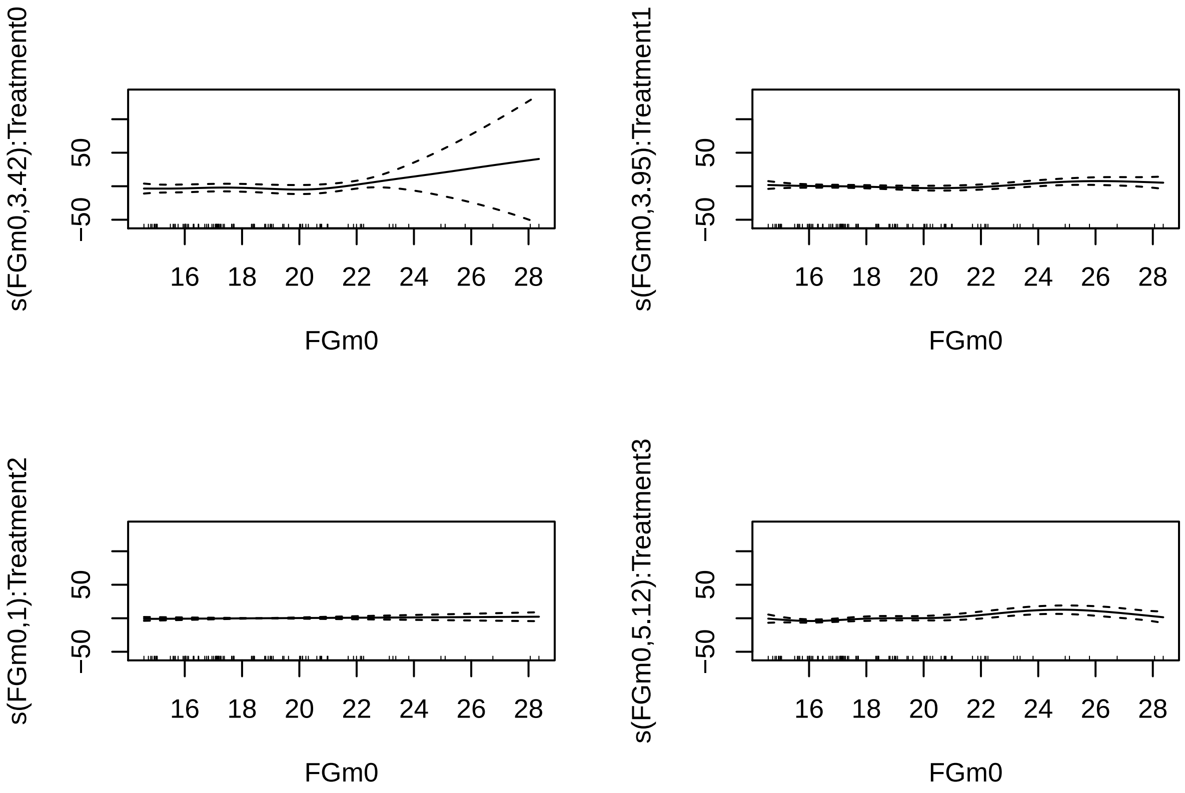


In this case the shape of the smooth terms for *FGm0* are different for each level of *Treatment*, with *Treatment0* and *Treatment2* being linear.

Next, we refit the model removing the non-significant terms, *height* and *DiaPres*.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + Treatment + SysPres + weight
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.53600    4.99808   3.909 0.000209 ***
## Treatment1  -6.98906    2.87166  -2.434 0.017440 *
## Treatment2  -6.58494    2.87579  -2.290 0.024987 *
## Treatment3  -5.68528    2.83957  -2.002 0.049060 *
## SysPres     -0.07947    0.04404  -1.804 0.075365 .
## weight       0.06232    0.03819   1.632 0.107119
## ---
## 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 3.416  4.137 0.715 0.682536
## s(FGm0):Treatment1 3.950  4.804 1.877 0.092364 .
## s(FGm0):Treatment2 1.000  1.000 0.472 0.494366
## s(FGm0):Treatment3 5.116  6.155 4.441 0.000639 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.333   Deviance explained =   47%
## GCV = 23.165  Scale est. = 18.205    n = 91
```



We see that the smooth term of *FGm0* for *Treatment0* has changed greatly, having a much higher *edf*. We compare with the previous model.

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0, by = Treatment) + Treatment + SysPres + s(DiaPres) +
##     weight + height
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + Treatment + SysPres + weight
##   Resid. Df Resid. Dev      Df Deviance      F Pr(>F)
## 1    67.228       1223
## 2    68.904       1302 -1.6763  -79.005 2.6896 0.0843 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The models are not significantly different. We keep the previous model that had a higher explained deviance.

Conclusions

After comparing different semiparametric models, we keep the model with the following formula:

- `FGm12 ~ s(FGm0, by= Treatment) + Treatment + SysPres + s(DiaPres) + weight + height`

which differenciates the smooth term of *FGm0* by *Treatment.* This model has an explained deviance percentage of 50.23% and an adjusted R^2 of 0.3582.

Lastly, we would like to highlight that there are few cases per *Treatment* level and that impacts the quality of the model obtained. With more data we would expect to get better models that do not differ so much between them.