# 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 we are going to work with is composed by 100 instances and 9 variables, it has missing values. All the variables are integers in exception of the treatment variable which is a factor.

| 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 |

The main objective of this project is to fit several GAM 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). Once generated the model, we are going to get an insight of how good each model is and finally, we will use function ANOVA to select among the best ones. To work comfortably with the models, before starting with the modelling part, we have removed the NA values from our dataset and we have converted the Treatment column as factors.

Aside from solving the NA's, we have performed a data exploration of the Treatment variable (using the script provided). In the aforementioned script, we have boxplots where we can see that FG decreases as the treatment has an effect on the patient so we can see a decrease between FG in the different stages (0,3,6,12). If we see FG between different treatments, we can see that the slope variates depending on the treatment that the patient is exposed, which means that one treatment or another will affect at the client hirsutism in a different way. We can see also that there are some treatments with a higher variance than others, for instance Treatment 2 has a very small variance whereas Treatment 0 has way more variance.

## Model 1

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

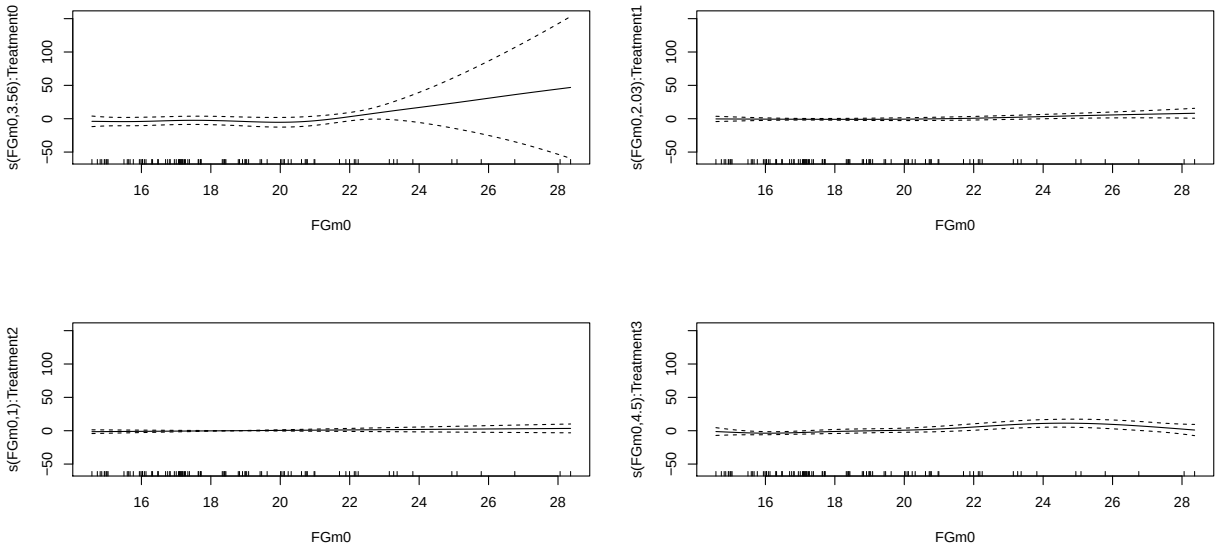
```
model.1 <- gam(FGm12 ~ weight + height + DiaPres + SysPres + FGm0 + Treatment, data=hirsutism)
summary(model.1)
#vis.gam(model.1)
```

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.

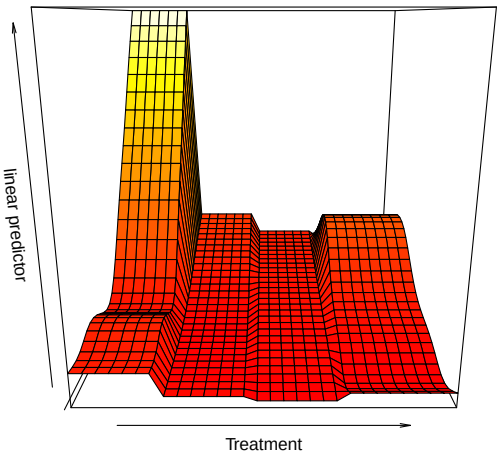
## Model 2

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




```
vis.gam(model.2)
```

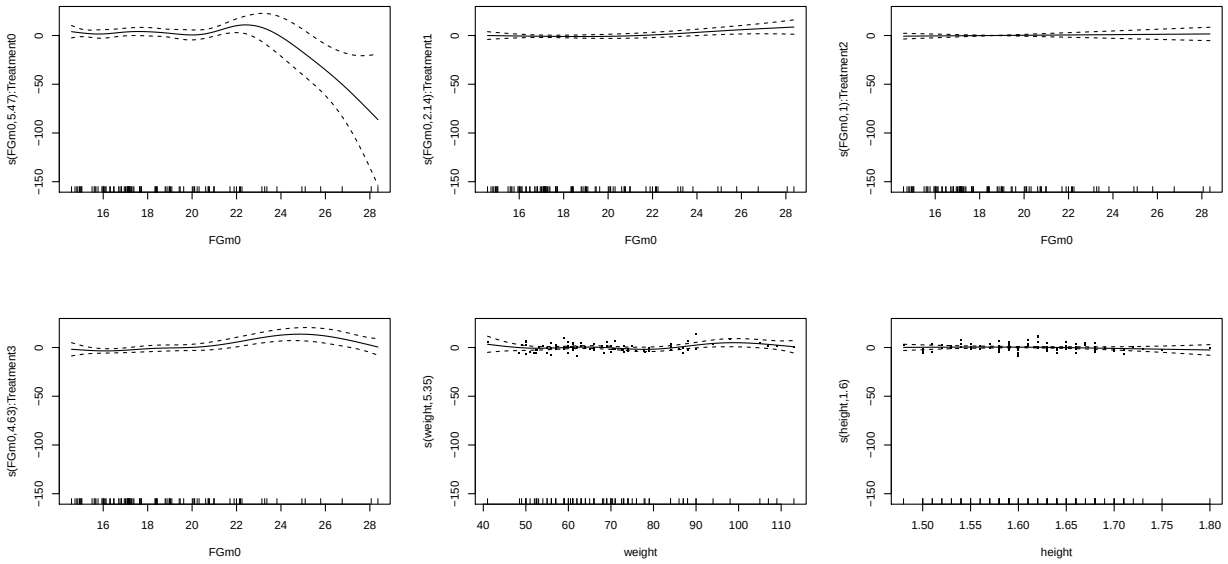



This model is greater than model 1 because R-sq. (adj) is about 0.291. We can see that Treatment 0 is the only one that does not seem to be linear while the others are.

## 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)
plot(model.3, pages=1, residuals=TRUE)
```

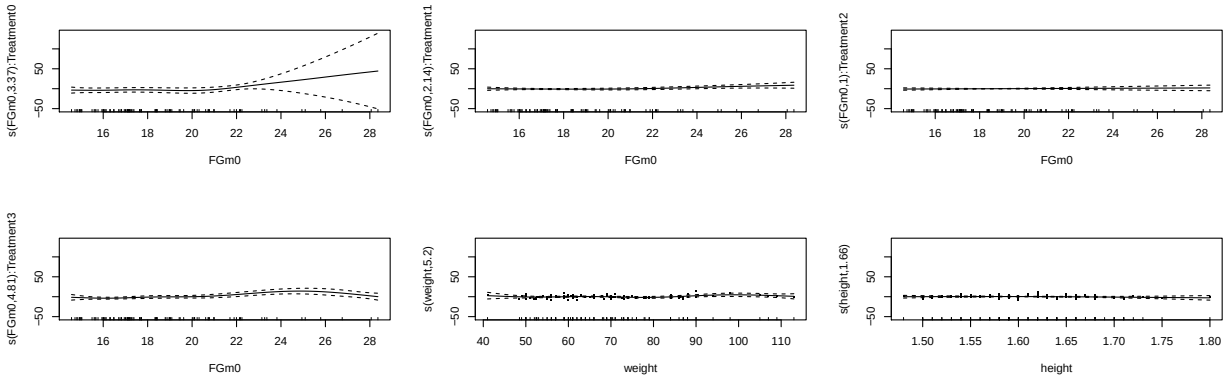



We can see again that height and weight are not relevant if we observe p-value. Otherwise, the R squared is about 0.367 and is better than the previous one.

## Model 4

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

```
model.4 <- gam(FGm12 ~ s(FGm0, by=Treatment) + s(weight) + s(height) + Treatment, data=hirsutism)
summary(model.4)
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")
summary(model.1)
summary(model.2)
```

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")
summary(model.2)
summary(model.3)
```

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")
summary(model.2)
summary(model.4)
```

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.