Gams

Assignment 2

Ignasi Mane, Antoni Company & Chiara Barbi

24/3/2019

Introduction

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. It refers to a male pattern of body hair (androgenic hair) and it is therefore primarily of cosmetic and psychological concern. Hirsutism is a symptom rather than a disease and may be a sign of a more serious medical condition, especially if it develops well after puberty.

The amount and location of the hair is measured by a Ferriman-Gallwey score. The original method used 11 body areas to assess hair growth, but was decreased to 9 body areas in the modified method: Upper lip, Chin, Chest, Upper back, Lower back, Upper abdomen, Lower abdomen, Upper arms, Thighs, Forearms (deleted in the modified method) and Legs (deleted in the modified method). In the modified method, hair growth is rated from 0 (no growth of terminal hair) to 4 (extensive hair growth) in each of the nine locations. A patient's score may therefore range from a minimum score of 0 to a maximum score of 36.

A clinical trial was conducted to evaluate the effectiveness of an antiandrogen combined with an oral contraceptive in reducing hirsutism for 12 consecutive months. It is known that contraceptives have positive effects on reduction of hirsutism. The degree of hirsutism is measured by the modified Ferriman-Gallwey scale. Patients were randomized into 4 treatment levels: levels 0 (only contraceptive), 1, 2, and 3 of the antiandrogen in the study (always in combination with the contraceptive). The clinical trial was double-blind.

The data set hirsutism.dat contains artifucial 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 grater 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.

Questions

1). 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). Use functions summary, plot and vis.gam to get an insight into the fitted models. Then use function anova to select among them the model (or models) that you think is (are) the most appropriate.

We load the required libraries

```
library(mgcv)
library(ggplot2)
library(emmeans)
```

We load the data stored in the file hirsutism.dat and use the function str to check the type of variables. We also erase those variables not needed for the assignment, that is, the variables that were not taken at the beginning of the experiment. To make things simpler, we will use the default tuning parameters that the function gam provides to us in the entire document

```
hirs <- read.table("hirsutism.dat",header=T, sep="\t",fill=TRUE)
hirs <- na.omit(hirs[,-c(3:4)])
str(hirs)
```

```
## 'data.frame':    91 obs. of  7 variables:
##  $ Treatment: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FGm0     : num  20.1 16.7 18.8 14.8 17.4 ...
##  $ FGm12    : num  6.57 9.47 16.11 11.86 4.18 ...
##  $ SysPres  : num  136 120 90 110 110 115 120 120 100 120 ...
##  $ DiaPres  : num  71 78 65 70 70 50 60 70 70 70 ...
##  $ weight   : num  86 52.4 63 64 62 ...
##  $ height   : num  1.71 1.58 1.63 1.57 1.6 ...
##  - attr(*, "na.action")= 'omit' Named int  8 28 36 42 43 53 60 83
##   ..- attr(*, "names")= chr  "8" "28" "36" "42" ...
```

As required, we define the variable Treatment as a factor with levels 0, 1, 2, and 3. Each level represents a type of treatment used to treat the disease.

```
hirs$Treatment <- factor(hirs$Treatment)
```

We take a look to the data using the function summary

```
summary(hirs)
```

```
##  Treatment      FGm0           FGm12           SysPres
##  0:22      Min.   :14.57   Min.   :-1.163   Min.   : 88.0
##  1:22      1st Qu.:16.40   1st Qu.: 5.566   1st Qu.:110.0
##  2:22      Median :17.70   Median : 8.069   Median :115.0
##  3:25      Mean   :18.67   Mean   : 9.053   Mean   :115.9
##            3rd Qu.:20.27   3rd Qu.:12.402   3rd Qu.:120.0
##            Max.   :28.36   Max.   :22.759   Max.   :162.0
##     DiaPres          weight          height
##  Min.   :46.00   Min.   : 41.00   Min.   :1.480
##  1st Qu.:65.00   1st Qu.: 57.00   1st Qu.:1.580
##  Median :70.00   Median : 64.00   Median :1.610
##  Mean   :70.04   Mean   : 68.06   Mean   :1.613
##  3rd Qu.:75.00   3rd Qu.: 74.50   3rd Qu.:1.650
##  Max.   :95.00   Max.   :113.00   Max.   :1.800
```

We compute the correlation matrix to see which covariates are linearly correlated.

```
cor <- round(cor(hirs[,2:7], use = "complete.obs"),2)
print(cor)
```

```
##         FGm0 FGm12 SysPres DiaPres weight height
## FGm0    1.00  0.31    0.01    0.09   0.22   0.11
## FGm12   0.31  1.00   -0.18   -0.08   0.00  -0.06
## SysPres 0.01 -0.18    1.00    0.64   0.45   0.18
## DiaPres 0.09 -0.08    0.64    1.00   0.47   0.20
## weight  0.22  0.00    0.45    0.47   1.00   0.38
## height  0.11 -0.06    0.18    0.20   0.38   1.00
```

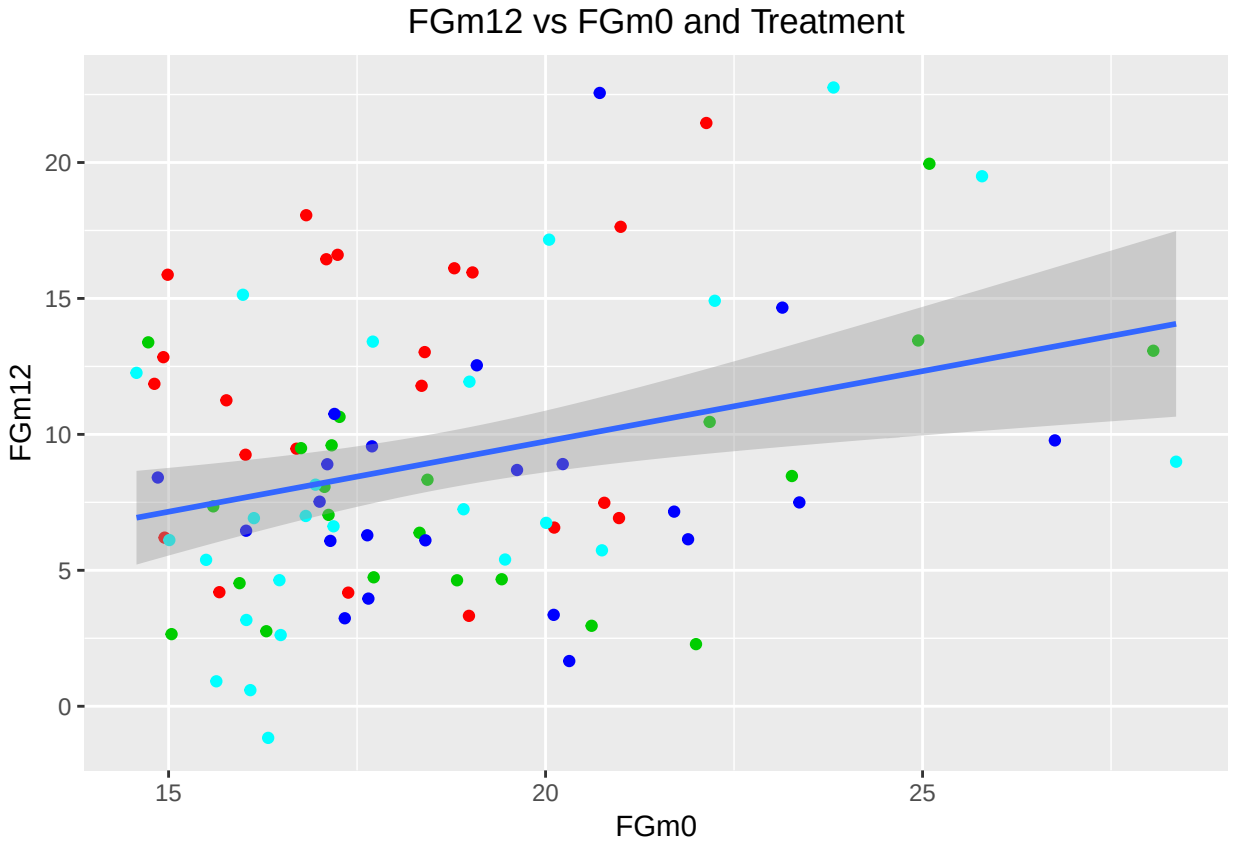
We look for variables with possible prediction power. As we see in the previous correlation table, it seems that FGm12 is poorly linearly correlated with its potential covariates. FGm0 might have something to say but in general no linear relation seems to exist between the response variable and the covariates.

Fitting a linear model

We will try to fit a linear model with two covariates: the factor treatment and the continuous variable FGm0. First we plot the data to detect any possible pastern in the data.

```
ggplot(hirs,aes(x=FGm0,y=FGm12))+
  geom_point(aes(FGm0,FGm12), col=as.numeric(hirs$Treatment)+1)+
  ggtitle(label="FGm12 vs FGm0 and Treatment")+
```

```
  theme(plot.title = element_text(hjust = 0.5))+
  geom_smooth(method = "lm")
```



We want to check whether there are differences between the four types of treatments. To do so we will fit a linear model. First doing only an ANOVA with one factor (Treatment) and then trying to add the covariate FGm0 to see whether it is significant.

```
mod.glm0 <- lm(FGm12~Treatment, data = hirs)
summary(mod.glm0)
```

```
##
## Call:
## lm(formula = FGm12 ~ Treatment, data = hirs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6494 -3.2995 -0.6682  3.0724 14.3659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.659      1.086  10.737   <2e-16 ***
## Treatment1    -3.708      1.536  -2.414   0.0179 *
```

```
## Treatment2    -3.466      1.536  -2.257   0.0265 *
## Treatment3    -3.173      1.489  -2.131   0.0359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.093 on 87 degrees of freedom
## Multiple R-squared:  0.08156,    Adjusted R-squared:  0.04989
## F-statistic: 2.575 on 3 and 87 DF,  p-value: 0.05901
```

```
mod.glm1 <- lm(FGm12~Treatment+FGm0, data = hirs)
summary(mod.glm1)
```

```
##
## Call:
## lm(formula = FGm12 ~ Treatment + FGm0, data = hirs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0922 -3.3090 -0.2387  3.0833 13.4913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5589     3.0695   0.182 0.855956
## 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 *
## FGm0          0.6247     0.1631   3.829 0.000244 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.735 on 86 degrees of freedom
## Multiple R-squared:  0.2153, Adjusted R-squared:  0.1788
## F-statistic:   5.9 on 4 and 86 DF,  p-value: 0.0003037
```

```
anova(mod.glm0,mod.glm1, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: FGm12 ~ Treatment
## Model 2: FGm12 ~ Treatment + FGm0
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     87 2256.7
## 2     86 1928.0  1    328.71 14.662 0.0002436 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we compare both models using the function anova, we can say that we do have statistical

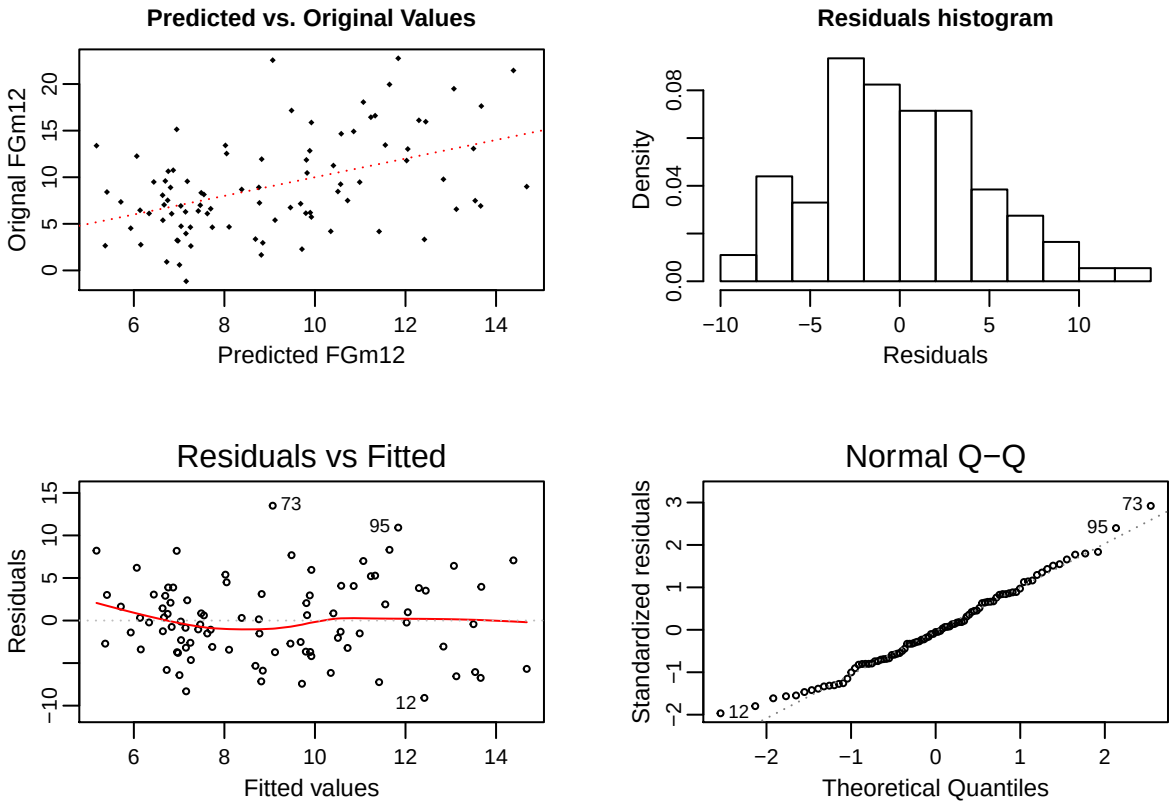
evidence to conclude that the covariate FGm0 is relevant to explain the data. Therefore, we chose the former model. To check whether there are differences between treatments we apply the tukey test using the package emmeans.

```
emm<-emmeans(mod.glm1,~Treatment)
cld(emm)
```

```
##  Treatment    emmean        SE df  lower.CL  upper.CL .group
##  1           7.636528 1.0128076 86  5.623133  9.649923  1
##  2           7.788233 1.0149744 86  5.770531  9.805936  1
##  3           8.623613 0.9476429 86  6.739761 10.507465  12
##  0          12.221824 1.0201166 86 10.193899 14.249748   2
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 4 estimates
## significance level used: alpha = 0.05
```

Applying this method we can conclude that treatment 0 is less effective than 1 and 2 but we do not have statistical evidence to say that treatment 3 and 0 have different effectiveness. We check the hypothesis of the model with classical diagnostics plots.

```
par(mfrow=c(2,2), mgp=c(1.5,0.5,0),oma=c(0,0,0,0),mar=c(4,3,2,2))

plot(x=mod.glm1$fitted.values,y=hirs$FGm12, pch=18, cex=0.5,
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9,
     xlab = "Predicted FGm12",
     ylab = "Orignal FGm12",
     main = "Predicted vs. Original Values")
abline(a=0,b=1, col=2, lty=3)

hist(x = mod.glm1$residuals, freq = FALSE,
     main = "Residuals histogram",
     xlab = "Residuals",breaks = 15,
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9)

plot(mod.glm1, which=c(1:2),
     cex=0.5, cex.main=0.9,cex.lab=0.9,cex.axis=0.8,
     sub.caption = "")
```



The residuals diagnostics seems to confirm the hypothesis of the linear model. Finally, we will use the function gam to create a gam object that will allow to compare this model to the ones that we will try later on this document. Notice that the model obtained is equivalent.

```
mod.glm1 <- gam(FGm12~Treatment+FGm0, data = hirs)
```

Fitting a semiparametric model 1

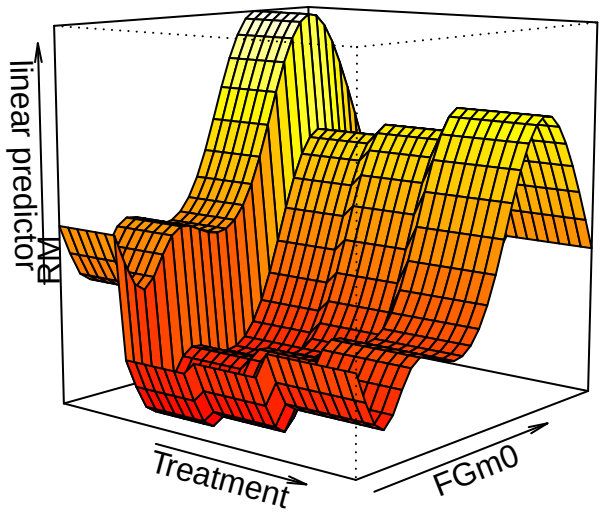
We fit the following semi parametric model:

```
mod.glm2 <- gam(FGm12~Treatment+s(FGm0), data = hirs)
summary(mod.glm2)
```

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

```
vis.gam(mod.glm2,se=0,theta =40, phi = 10, d=4,nticks=3)
text(-.61,-.1,'RM',srt=90)
```



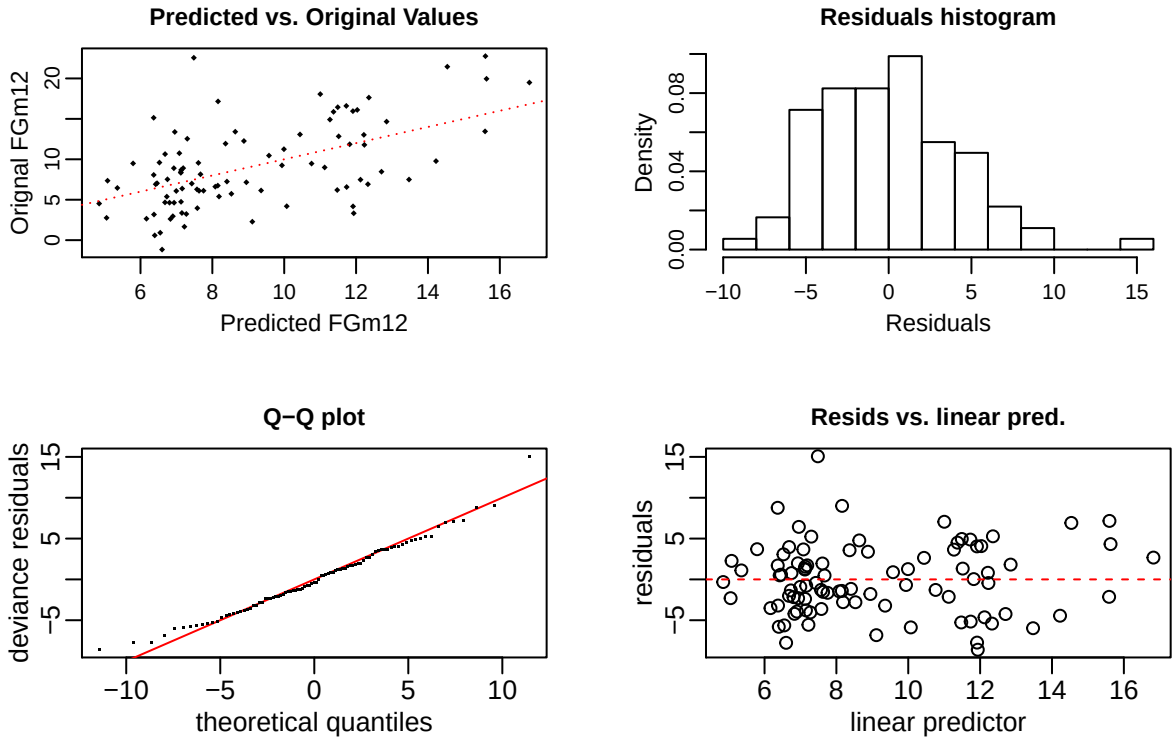
Once we have summarized the model, we can see that all the parametric coefficients analyzed are significant with a p-value smaller than 0.05 and, for the non-parametric, FGm0 is significant with a p-value lower than 0.05.

We can also see that the adjustment of the model is really low with a value of $R^2 = 0.259$ and with a deviance explained of only the 33.1%.

If we analyse the residuals of the model

```
par(mfrow=c(2,2), mgp=c(1.5,0.5,0),oma=c(1,0,1,0),mar=c(4,3,2,2))

plot(x=mod.glm2$fitted.values,y=hirs$FGm12, pch=18, cex=0.5,
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9,
     xlab = "Predicted FGm12",
     ylab = "Orignal FGm12",
```

```
     main = "Predicted vs. Original Values")
abline(a=0,b=1, col=2, lty=3)

hist(x = residuals(mod.glm2), freq = FALSE,
     main = "Residuals histogram",
     xlab = "Residuals",breaks = 15,
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9)

qq.gam(mod.glm2, rep = 0, level = 0.9, rl.col = 2,
       rep.col = "gray80",main="Q-Q plot",cex.main=0.9)

plot(napredict(mod.glm2$na.action, mod.glm2$linear.predictors),
     residuals(mod.glm2), main = "Resids vs. linear pred.",
     xlab = "linear predictor", ylab = "residuals",cex.main=0.9)
abline(h=0, lty=2, col="red")
```



Fitting a semiparametric model 2

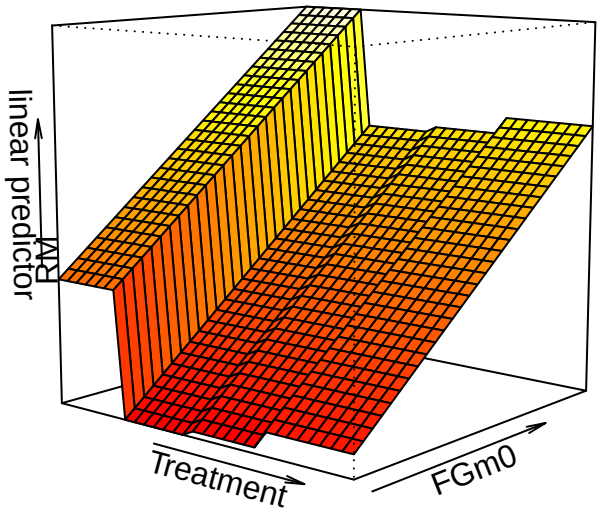
We fit the following semi parametric model:

```
mod.glm4 <- gam(FGm12~Treatment+s(FGm0, SysPres), data = hirs, na.action = na.omit)
summary(mod.glm4)
```

```
##
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ Treatment + s(FGm0, SysPres)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.062      1.022  11.797  < 2e-16 ***
## Treatment1    -4.312      1.454  -2.967  0.00391 **
## Treatment2    -4.117      1.462  -2.815  0.00606 **
## Treatment3    -3.535      1.383  -2.556  0.01237 *
## ---
## 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,SysPres)   2      2 8.298 0.000499 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.186   Deviance explained = 23.2%
## GCV = 23.781  Scale est. = 22.213    n = 91
```

```
vis.gam(mod.glm4,se=0,theta =40, phi = 10, d=4,nticks=3)
text(-.61,-.1,'RM',srt=90)
```

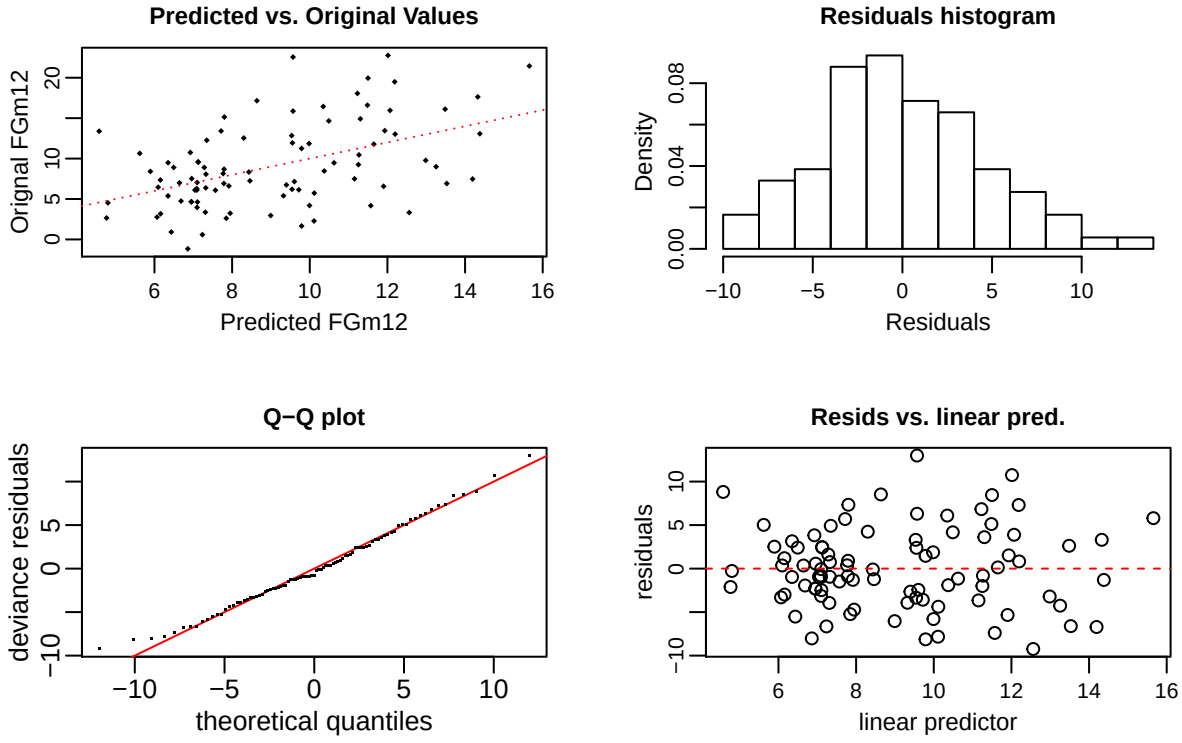


If we analyse the results of the model, we can see that all the parametric coefficients analyzed are significant with a p-value smaller than 0.05 and also for the non-parametric ones.

We can also see that the adjustment of the model is extremely low with a value of $R^2 = 0.186$ and with a deviance explained of only the 23.2%.

If we proceed to analyse the residuals of the model

```
par(mfrow=c(2,2), mgp=c(1.5,0.5,0),oma=c(1,0,1,0),mar=c(4,3,2,2))

plot(x=mod.glm4$fitted.values,y=hirs$FGm12, pch=18, cex=0.5,
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9,
     xlab = "Predicted FGm12",
     ylab = "Orignal FGm12",
     main = "Predicted vs. Original Values")
abline(a=0,b=1, col=2, lty=3)

hist(x = residuals(mod.glm4), freq = FALSE,
     main = "Residuals histogram",
     xlab = "Residuals",breaks = 15,
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9)

qq.gam(mod.glm4, rep = 0, level = 0.9, rl.col = 2,
       rep.col = "gray80", main="Q-Q plot",cex.main=0.9)

plot(napredict(mod.glm4$na.action, mod.glm4$linear.predictors),
     residuals(mod.glm4), main = "Resids vs. linear pred.",
     xlab = "linear predictor", ylab = "residuals",
     cex.lab=0.9,cex.axis=0.8,cex.main=0.9)
abline(h=0, lty=2, col="red")
```



Model Comparaison

Finally we will use to criteria to compare the three models that we defined. First we will use the function anova to compare the models pairwise using the Chi square test and then we will apply the Akaike Information Criteria.

```
anova(mod.glm1,mod.glm2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ Treatment + FGm0
## Model 2: FGm12 ~ Treatment + s(FGm0)
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    86.000     1928.0
## 2    80.108     1643.8 5.8921   284.13  0.02736 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.glm4,mod.glm2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ Treatment + s(FGm0, SysPres)
## Model 2: FGm12 ~ Treatment + s(FGm0)
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    85.000     1888.1
## 2    80.108     1643.8 4.8921   244.22   0.0316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to this criteria, which does not penalize the complexity of the model, the best model is "mod.glm2"

```
AIC(mod.glm1,mod.glm2,mod.glm4)
```

```
##                df      AIC
## mod.glm1  6.00000 548.1038
## mod.glm2 10.76304 543.1214
## mod.glm4  7.00000 548.1999
```

According to this criteria, which does penalize the complexity of the model, the best model is also "mod.glm2"