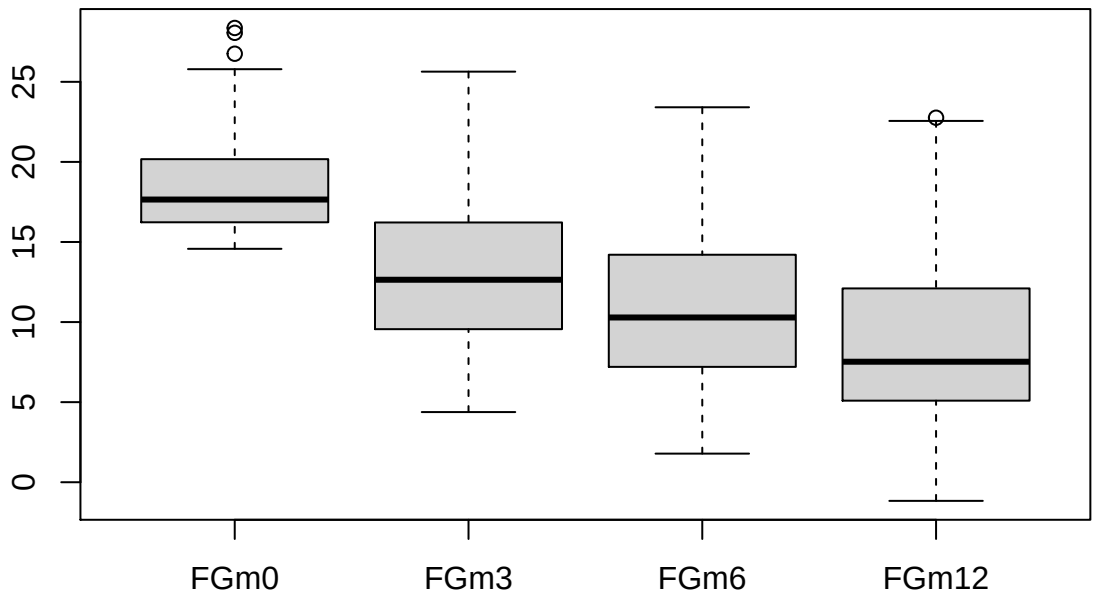
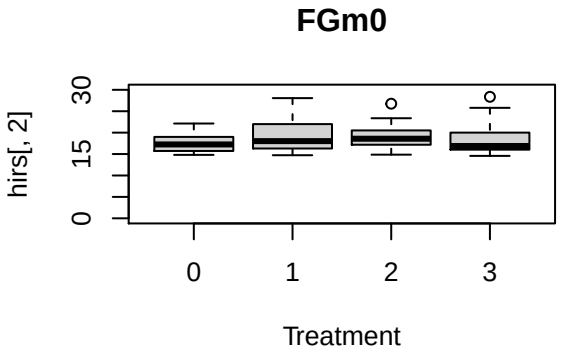
GAMs for hirsutism data

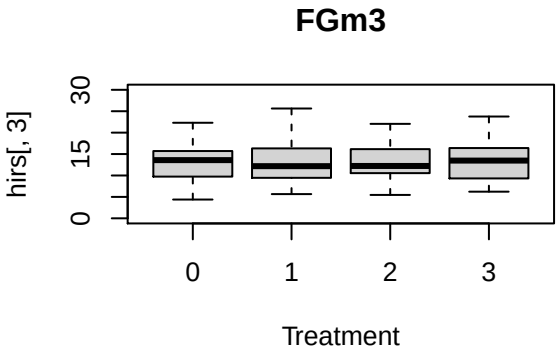
Geraldo Gariza, Jan Leyva and Andreu Meca
Universitat Politècnica de Catalunya

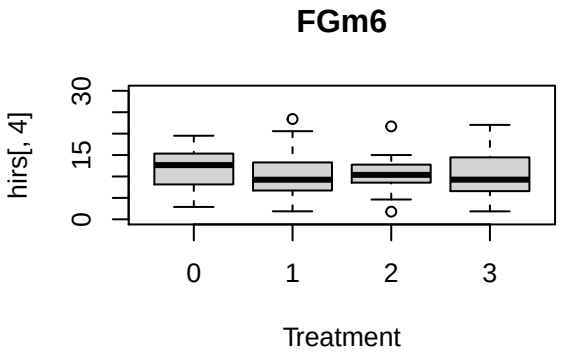
4th of April, 2021

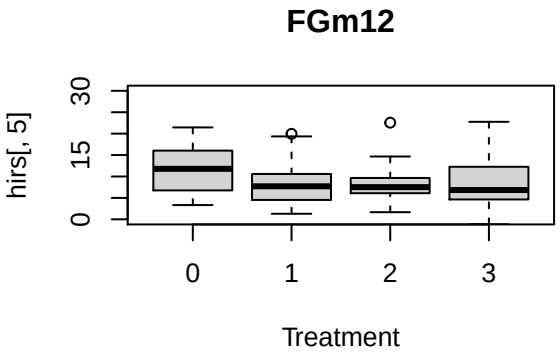
Fit several GAM models (including semiparametric models) explaining FGm12 as a function of the variables that were measured at the beginning of the clinical trial (including FGm0) and Treatment (treated as factor). Use functions summary, plot and vis.gamto get an insight into the fitted models. Then use function anova to select among them the model (or models) that you think is (are) the most appropriate.

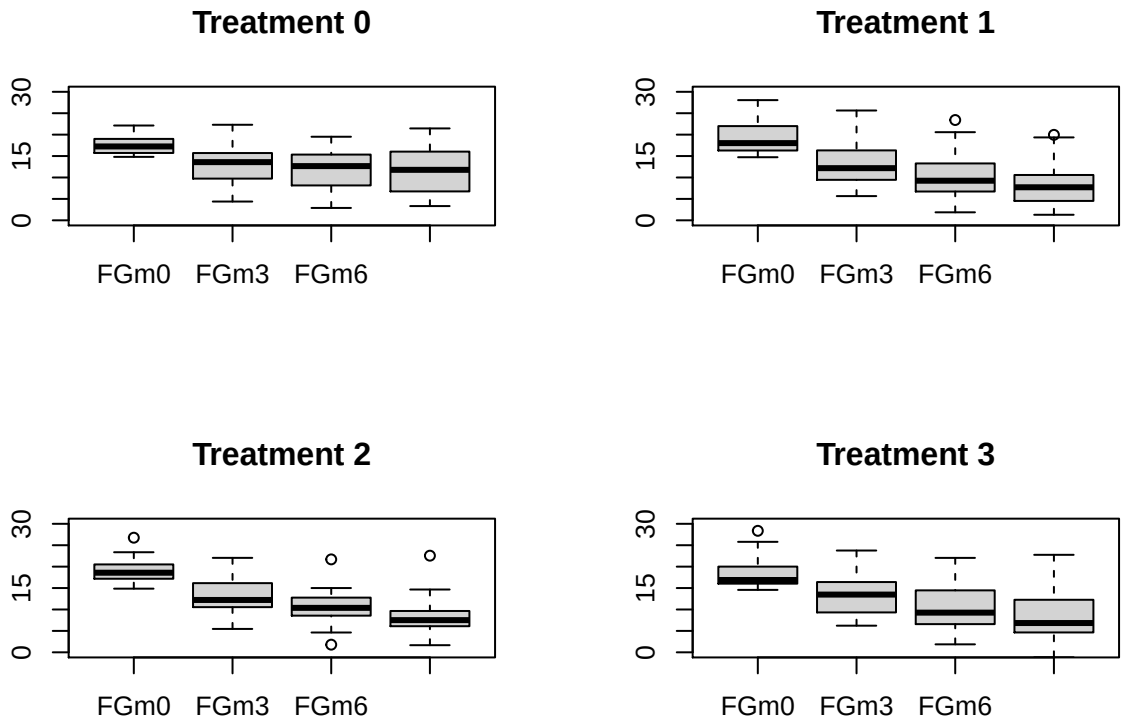



FGm0
hirs[, 2]
0
15
30
0
1
2
3
Treatment


FGm3
hirs[, 3]
0
15
30
0
1
2
3
Treatment


FGm6
hirs[, 4]
0
15
30
0
1
2
3
Treatment


FGm12
hirs[, 5]
0
15
30
0
1
2
3
Treatment




```
## GAM models: ------------------------------------------------------------

# Simples model: ---------------------------------------------------------
gam_simple <- gam(FGm12 ~ (FGm0) + (Treatment) + (SysPres)
             + (DiaPres) + (weight) + (height),
             data = hirs)

summary(gam_simple) # 19 %
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ (FGm0) + (Treatment) + (SysPres) + (DiaPres) + (weight) +
##     (height)
##
```

```
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.26663   14.82851   1.771  0.08013 .
## FGm0          0.52999    0.16948   3.127  0.00243 **
## Treatment    -1.16799    0.47107  -2.479  0.01516 *
## SysPres      -0.08608    0.05285  -1.629  0.10712
## DiaPres       0.01364    0.07207   0.189  0.85033
## weight        0.04081    0.04475   0.912  0.36446
## height      -11.81451    9.16867  -1.289  0.20108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## R-sq.(adj) =  0.134   Deviance explained = 19.2%
## GCV = 25.604  Scale est. = 23.634    n = 91
```

```
# gam.check(gam_simple)
```

Using the simplest model, GAM function explaining $FGm12$ as function of: $(FGm0) + (Treatment) + (SysPres) + (DiaPres) + (weight) + (height)$, we obtain only 19% of deviance explained. We should make some adjustment to the model.

Also we see some irregularities in the Q-Q plot of residuals. The cues have some deviations.

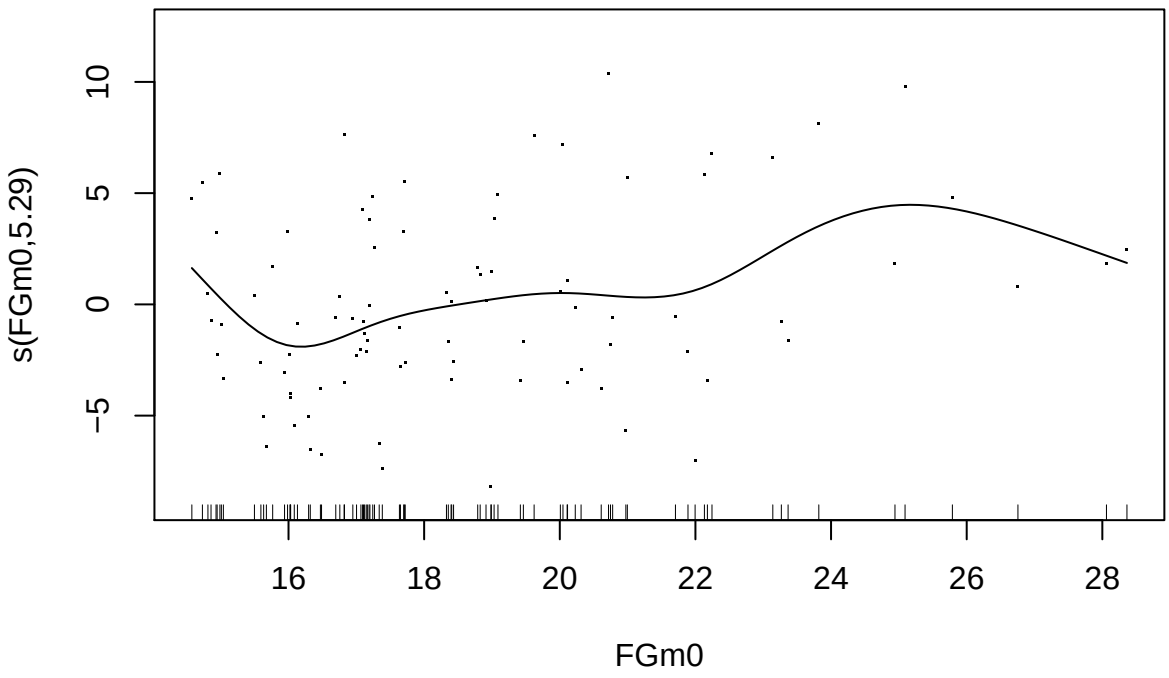
```
# Simple model 1: ------------------------------------------------------
gam_1 <- gam(FGm12 ~ s(FGm0) + (Treatment) + s(SysPres)
          + s(DiaPres) + s(weight) + s(height),
            data = hirs)
summary(gam_1) # 47 %
```

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

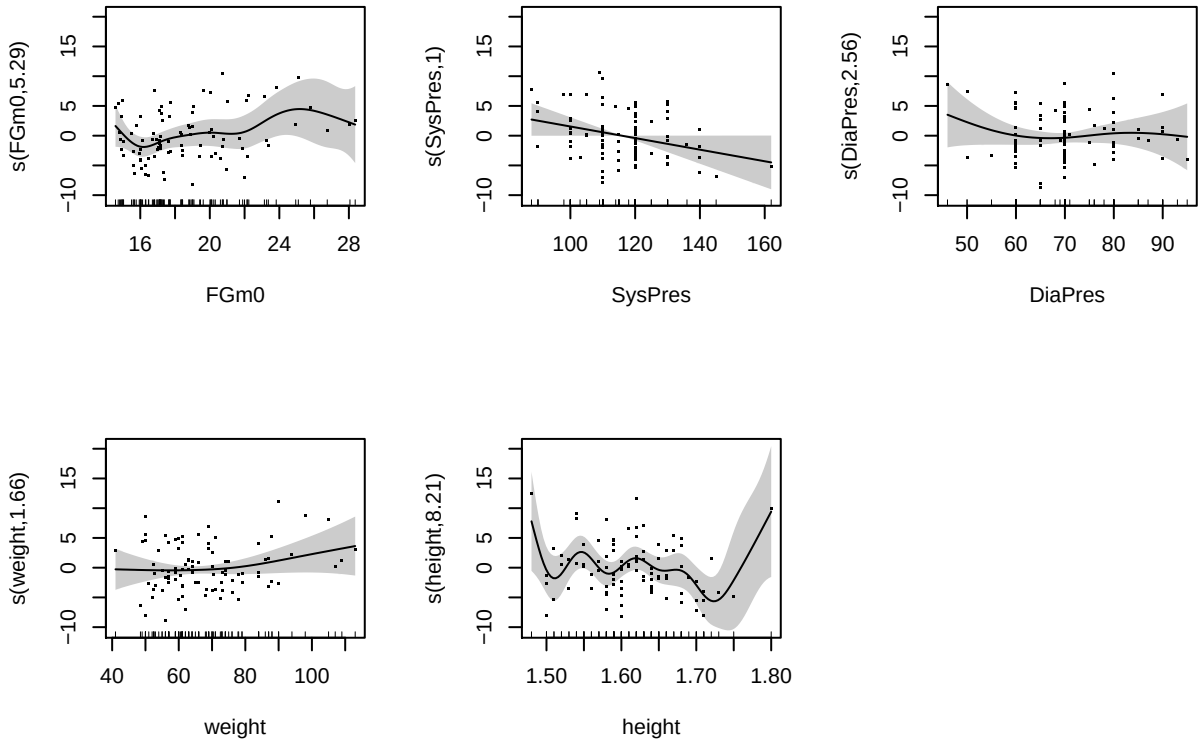
```
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5791     0.8447  13.708  < 2e-16 ***
## Treatment    -1.6304     0.4608  -3.538  0.00072 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(FGm0)    5.289  6.324 1.187  0.2992
## s(SysPres) 1.000  1.000 3.963  0.0504 .
## s(DiaPres) 2.562  3.211 0.703  0.5457
## s(weight)  1.659  2.057 1.247  0.2973
## s(height)  8.212  8.778 1.820  0.0678 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.321   Deviance explained =   47%
## GCV = 23.998  Scale est. = 18.533    n = 91
```

```r
# Plot

plot(gam_1, select = 1, residuals = TRUE, se=FALSE)
```




```
plot(gam_1,pages=1,residuals=TRUE,scheme=TRUE)
```




```
# par(mfrow=c(2,2))
# plot(gam_1, residuals = TRUE, shade=TRUE, seWithMean=TRUE, pages = 2)
# par(mfrow=c(1,1))
```

Applying the `s()` function to the variables, except *Treatment*, we see a higher deviance explained: 47%. `s()` is referring to which explanatory variable we apply smoothing.

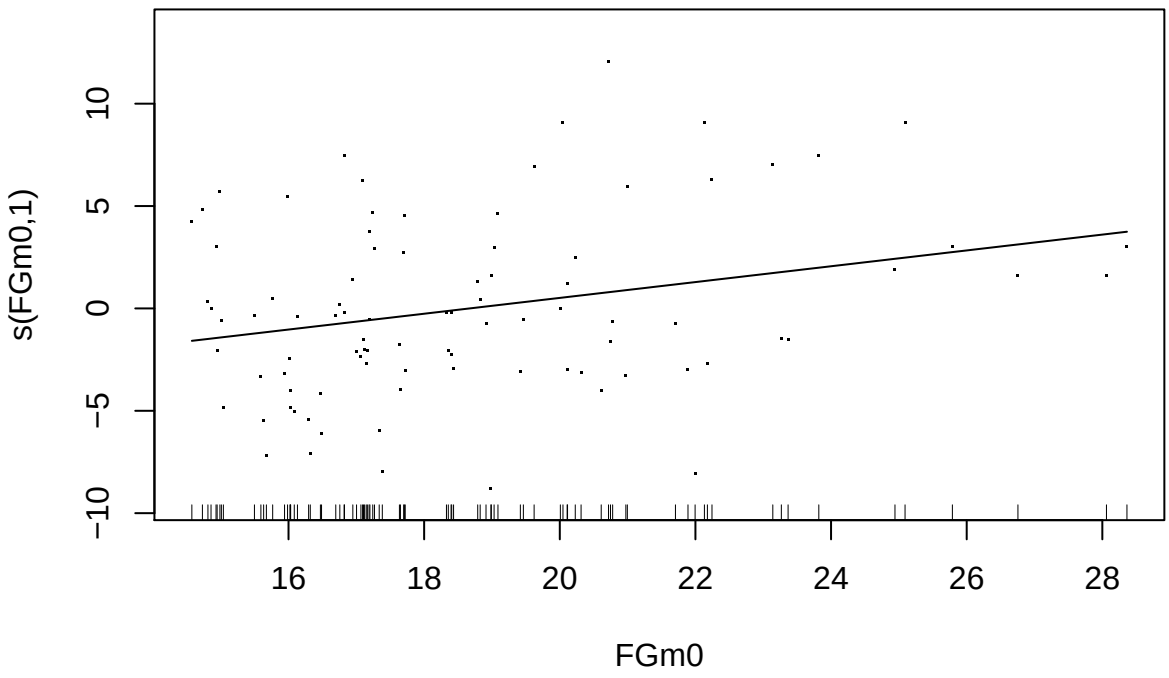
```
# Simple model 2: ------------------------------------------------------

# 'DiaPres' and 'weight' have a large p-value so can be removed from the model.
gam_2 <- gam(FGm12 ~ s(FGm0) + (Treatment) +
               s(SysPres) + s(height),
             data = hirs)
summary(gam_2) # 35.9 %
```

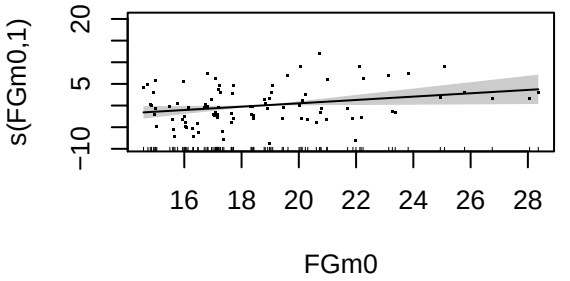
```
##
## Family: gaussian
```

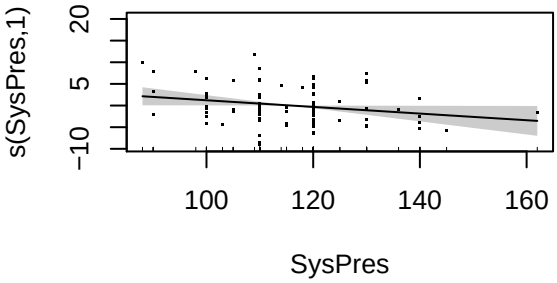
```
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.6296     0.8441  13.777  < 2e-16 ***
## Treatment    -1.6629     0.4530  -3.671 0.000438 ***
## ---
## 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)    1.000  1.000 4.890  0.0299 *
## s(SysPres) 1.000  1.000 4.239  0.0428 *
## s(height)  8.344  8.878 2.403  0.0257 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.267   Deviance explained = 35.9%
## GCV = 23.154  Scale est. = 20.013    n = 91
```

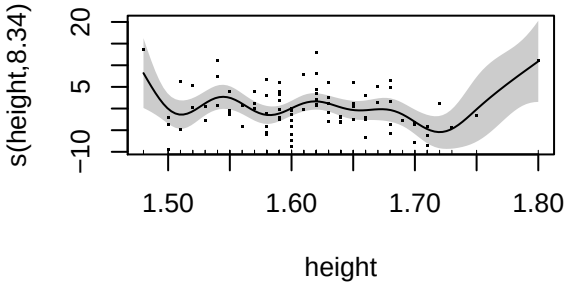
```r
plot(gam_2, select = 1, residuals = TRUE, se=FALSE)
```




```
plot(gam_2,pages=1,residuals=TRUE,scheme=TRUE)
```








```
# plot(gam_2, residuals = TRUE, shade=TRUE, seWithMean=TRUE, pages = 3)
```

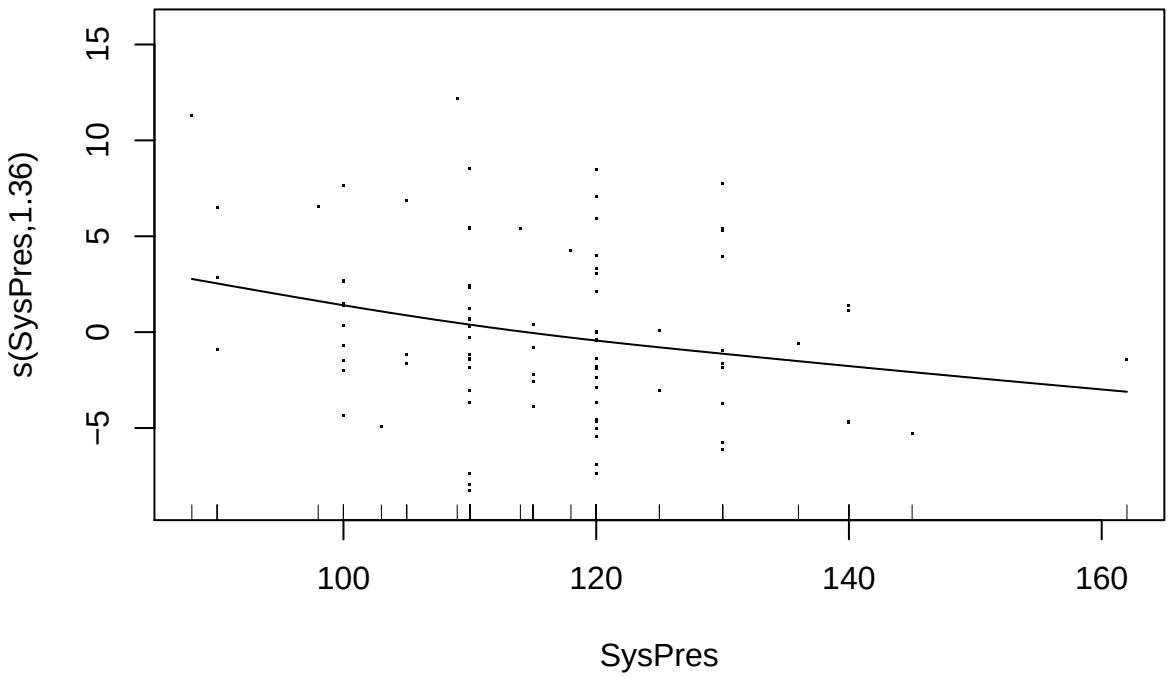
In model 2, we have removed two explanatory variables as they were non-significant, *DiaPres* and *weight*, with $p.value > 0.05$. Now, we see that all variables are significant but also the deviance explained of the model have decreased: 35,9%.

```
# Simple model 3: ------------------------------------------------------
gam_3 <- gam(FGm12 ~ (Treatment) +
               s(SysPres) + s(height),
             data = hirs)
summary(gam_3) # 33.4 %
```

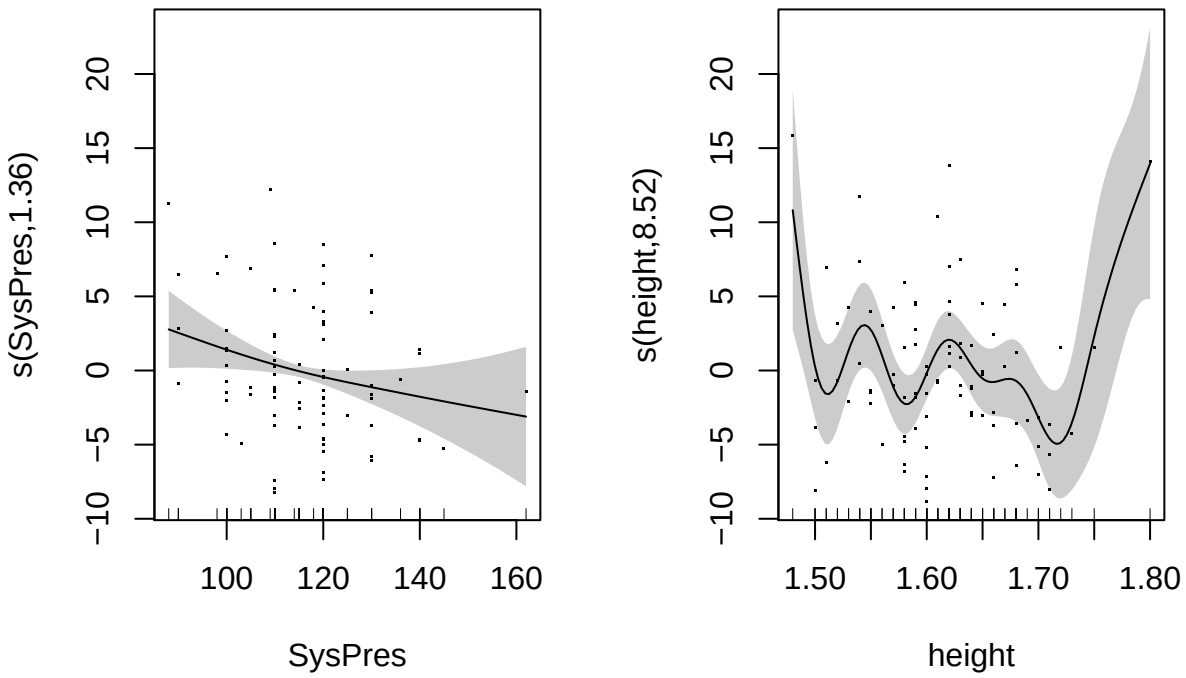
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

```
## FGm12 ~ (Treatment) + s(SysPres) + s(height)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.7520     0.8589  13.682  < 2e-16 ***
## Treatment    -1.7420     0.4611  -3.778 0.000305 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df     F p-value
## s(SysPres) 1.356  1.635 2.441 0.07052 .
## s(height)  8.522  8.932 3.272 0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.243   Deviance explained = 33.4%
## GCV = 23.783  Scale est. = 20.679    n = 91
```

```
plot(gam_3, select = 1, residuals = TRUE, se=FALSE)
```




```
plot(gam_3,pages=1,residuals=TRUE,scheme=TRUE)
```




```
# plot(gam_3, residuals = TRUE, shade=TRUE, seWithMean=TRUE, pages = 7)
```

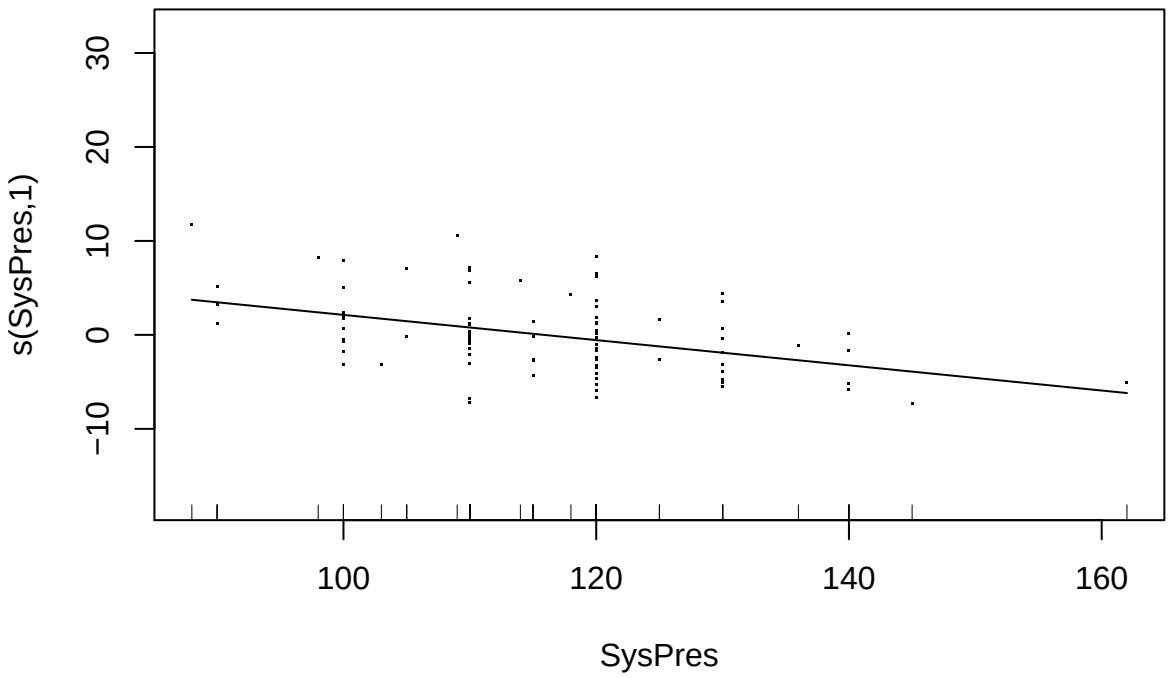
In the third model we are trying to see if removing the initial variable $FGm0$ we get better results. We would to show if we can explain the end result of the treatment without counting the initial level of hirsutism. We got slightly worse results with this model: 33,4%.

```
# model 4: ------------------------------------------------------------
gam_4 <- gam(FGm12 ~ (Treatment) +
               s(SysPres) + s(height) + te(height,weight),
             data = hirs)
summary(gam_4) # 46.1 %
```

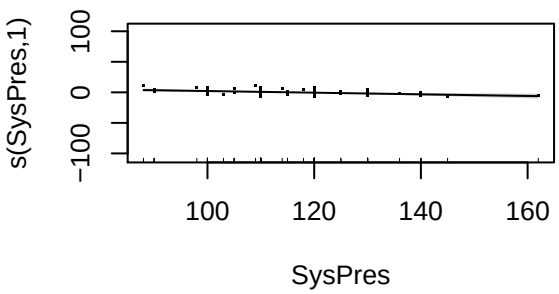
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
```

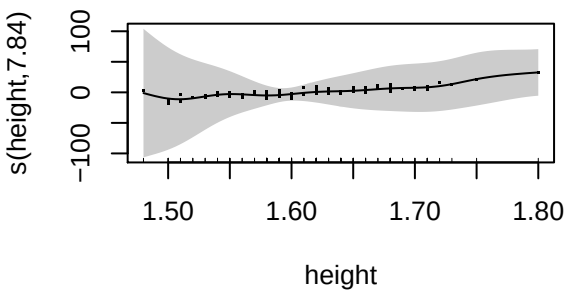
```
## FGm12 ~ (Treatment) + s(SysPres) + s(height) + te(height, weight)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.9513     0.8398  14.232  < 2e-16 ***
## Treatment    -1.8706     0.4580  -4.084 0.000113 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df     F p-value
## s(SysPres)        1.000  1.000 9.797 0.00252 **
## s(height)         7.841  8.378 1.010 0.29685
## te(height,weight) 7.994  9.562 0.913 0.58367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.328   Deviance explained = 46.1%
## GCV = 23.138  Scale est. = 18.349    n = 91
```

```
plot(gam_4, select = 1, residuals = TRUE, se=FALSE)
```




```
plot(gam_4,pages=1,residuals=TRUE,scheme=TRUE)
```



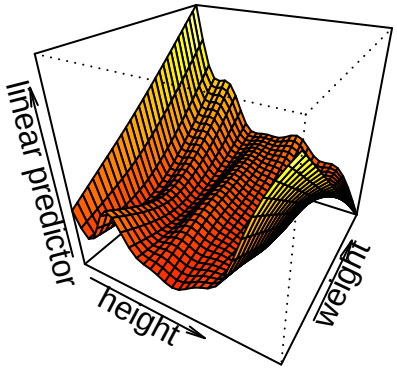


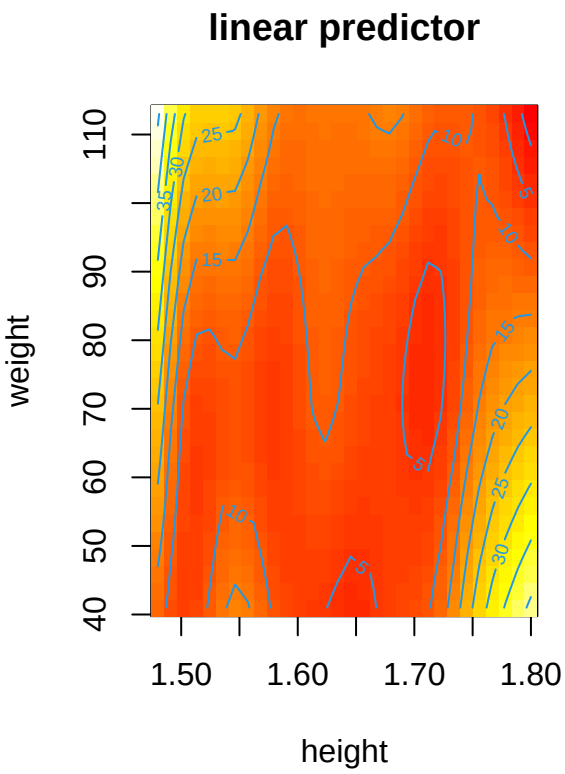

```
# plot(gam_4, residuals = TRUE, shade=TRUE, seWithMean=TRUE, pages = 7)
```

In this last model we use `te()` function to *height* and *weight*. We use the tensor product smooth as this two variables tend to be correlated. The result is very satisfying: 46,1%.

```
# We are going to visualize the joint effect of the variables:

par(mfrow=c(1,2))
vis.gam(gam_4, view=c("height","weight"), plot.type = "persp", theta=30, phi=30)
vis.gam(gam_4, view=c("height","weight"), plot.type = "contour")
```






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

- **ANOVA:**

Now we test the null hypothesis that states the `gam_simple` is correct again the alternative that states that the `gam_1` model is better:

```
anova(gam_simple, gam_1, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ (FGm0) + (Treatment) + (SysPres) + (DiaPres) + (weight) +
##     (height)
## Model 2: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(DiaPres) + s(weight) +
##     s(height)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    84.000     1985.3
## 2    67.629     1302.5 16.371    682.8 2.2505 0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis that `gam_simple` is better model than `gam_1`.

Now we test the null hypothesis that states the second model is correct again the alternative that states that the gam_1 is better:

```
anova(gam_2, gam_1 , test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(height)
## Model 2: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(DiaPres) + s(weight) +
##     s(height)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    78.122     1574.1
## 2    67.629     1302.5 10.492   271.67 1.3971 0.1977
```

We cannot reject h0, we do not have enough information, as the p-value > 0.05.

```r
anova(gam_3, gam_1, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ (Treatment) + s(SysPres) + s(height)
## Model 2: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(DiaPres) + s(weight) +
##     s(height)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    78.433     1636.2
## 2    67.629     1302.5 10.803   333.68 1.6666 0.1016
```

```r
anova(gam_4, gam_1, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ (Treatment) + s(SysPres) + s(height) + te(height, weight)
## Model 2: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(DiaPres) + s(weight) +
##     s(height)
##   Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
## 1    70.061     1324.2
## 2    67.629     1302.5 2.4312   21.696 0.4815 0.6568
```

```r
anova(gam_3, gam_2, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ (Treatment) + s(SysPres) + s(height)
## Model 2: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(height)
##   Resid. Df Resid. Dev      Df Deviance      F  Pr(>F)
## 1    78.433     1636.2
## 2    78.122     1574.1 0.31091   62.014 9.9664 0.01872 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing model 2 and 3, being h0 model 3 fits better the data, we reject that with p-value at 5%. We continue comparing gam_2.

```r
anova(gam_2, gam_4, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0) + (Treatment) + s(SysPres) + s(height)
## Model 2: FGm12 ~ (Treatment) + s(SysPres) + s(height) + te(height, weight)
##   Resid. Df Resid. Dev     Df Deviance    F Pr(>F)
## 1    78.122     1574.1
## 2    70.061     1324.2 8.0612   249.97 1.69 0.1156
```

```r
anova(gam_3, gam_4, test = 'F')
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ (Treatment) + s(SysPres) + s(height)
## Model 2: FGm12 ~ (Treatment) + s(SysPres) + s(height) + te(height, weight)
##   Resid. Df Resid. Dev     Df Deviance      F  Pr(>F)
## 1    78.433     1636.2
## 2    70.061     1324.2 8.3721   311.99 2.0309 0.05238 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We cannot choose between models 1,2,4. We can reject model 3 as ANOVA test was not significant against the other models. With the three remained We choose model 4 because of better explanation of the model with R-adjust 0.328 and explained variance 46.1%.