Generalized Additive Models for Hirsutism Data

Marcel Pons Cloquells

January 5, 2021

1 GAM Fit for Hirsutism Data

In this homework, data about the hirsutism condition in women and its treatments will be analyzed. Hirsutism is the excessive hairiness on women in those parts of the body where terminal hair does not normally occur or is minimal.

The degree of this condition is measured by a Ferriman Galley (FGm) score, which ranges from a minumum of 0 to a maximum of 36. The `Hirsutim_data` contains artificial values of 99[1] records of hirsutim, each one with 4 FG score measures that correspond to the degree of the condition along 12 months (i.e., baseline *FGm0*, *FGm4*, *FGm8*, *FGm12*) on which the patients are treated with anti-androgen combined with an oral contraceptive. There are 4 treatment levels into which patients were randomized: levels 0 (only contraceptive), 1, 2, and 3 of the antiandrogen in the study (always in combination with the contraceptive).

Apart from the variables defining the different FG measures and the treatment, the `Hirsutism_data` also contains baseline records of systolic and diastolic preassure and records of heigth and weight.

1.1 Generalized Additive Models

We will 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: *FGm0*, *Treatment*, *SysPres*, *DiaPres*, *Height* and *Weight*.

As a reference in order to compare results with other lessons done in the course, we have built a multiple linear regression model using all the aforementioned explanatory variables, which performs with a *Deviance* of 24.4% and an *Adjusted* R^2 of 0.17. In this model the only variables that result significant (p-value < 0.05) are *FGm0* and *Treatment*.

Model 1

The first `gam` model proposed includes all the explanatory variables estimated with non parametric smooth functions `s()` and with the *Treatment* factor variable.

```
m1 <- gam(FGm12 ~ s(FGm0) + s(SysPres) + s(DiaPres) + s(weight) + s(height) +␣
 ↪Treatment)
```

[1] We have dropped 8 rows which contained missing data, hence finally working with 91 records

```
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 0.00143 **
s(SysPres) 1.000  1.000 3.015 0.08653 .
s(DiaPres) 1.956  2.464 1.191 0.40264
s(weight)  1.000  1.000 1.146 0.28770
s(height)  1.000  1.000 1.454 0.23159
---
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
```

The model has a better *Deviance explained* (38.6%) and *Adjusted* R^2 (0.289) than the full linear model[2]. The variables that happen to be significant for the model are *Treatment* and *FGm0*. The other continuous variables have *Equivalent Degrees of Freedom* equal to 1, hence they seem to not need any smoothing function to increase their flexibility.

Model 2

For the second model we consider not applying `s()` to the variables that happen to have 1 (or approx.) *edf* in the model 1.

```
m2 <- gam(FGm12 ~ s(FGm0) + Treatment + SysPres + DiaPres + weight + height)
```

For this semiparametric model the *Deviance explained* and adj-R^2 decreases to 36.6% and 0.262, with *FGm0* and *Treatment* being the only significant variables as in the previous gam.

If we compare the two models already build with `Anova(m1,m2, test="F")` we accept the null hypothesis stating that the complex model `m1` does not differ significantly (p-value $= 0.1197$) from the simpler model `m2`. Therefore, following the law of parsimony, we should keep the later.

Model 3

For this model we consider excluding those variables that did not have a `p-value` lower than 0.05.

```
m3 <- gam(FGm12 ~ s(FGm0) + Treatment)
```

[2]Comparison Anova test yields a p-value of 0.01792 supporting as better the non parametric model (see .Rmd).

```
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

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(FGm0) 5.763  6.892 3.999 0.00102

R-sq.(adj) =  0.259   Deviance explained = 33.1%
GCV = 22.667  Scale est. = 20.235    n = 91
```

The model has lower *Deviance explained* and adj-R^2 with respect to the two previous models and the two considered variables result significant. Comparing with `Anova(m2,m3, test="F")` the null hypothesis is accepted (p-value $= 0.366$), keeping this model instead of `m2` which considers all the variables.

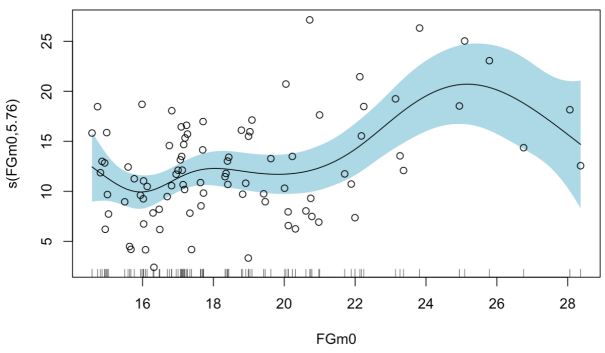
Model 4

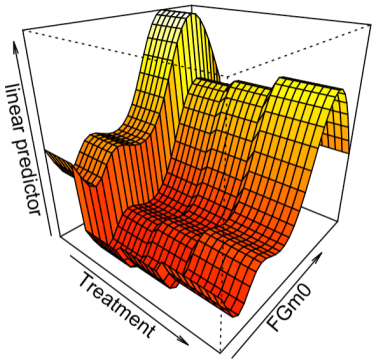
In the last model that we propose, we estimate the non parametric functions for each level of the *Treatment* factor variable, that is, we will have 4 estimators, each one for a level of the variable (factor-smooth interaction).

```
m4 <- gam(FGm12 ~ s(FGm0, by=Treatment))
```

Both *Deviance explained* and *Adjusted* R^2 are greater than `model 3`. If we compare both models with `Anova()`, we see that adding complexity to the model does not imply a significant difference in *Deviance* and *adj-*R^2 (p-value $= 0.1972$). Therefore, we keep the simpler `model 3` as our final model.

Visualization of model 3





From the left hand plot[3] it can be appreciated that in general the treatments are effective since the estimated values of FGm12 depicted on the y-axis are lower than the intital values of FGm0.

Furthermore, the FG measure at month 12 is greater for those patients who had a higher initial FG measure, therefore it is possible that the treatments have a limited effect and cannot decrease the hair of patients with different intial FG score to the same levels (e.g., the final FG score between a patient of 25 intial score and a patient of 16 initial score is improbable to be close).

From the right hand side plot we can visually analyze the effects of the different treatments[4]. Treatment 0 (only contraceptive) has values of linear predictor (i.e., FGm12) greater than the other treatments, therefore being the most ineffective treatment for hirsutirm. On the other hand, treatments 1 and 2 seem to be the ones that have a more effective outcome against the condition.

[3]To make the plot more interpretable, we shifted the scale so that the intercept is included with `shift = coef(model3)[1]`)

[4]The effects of the intial FGm0 score also can be appreciated.