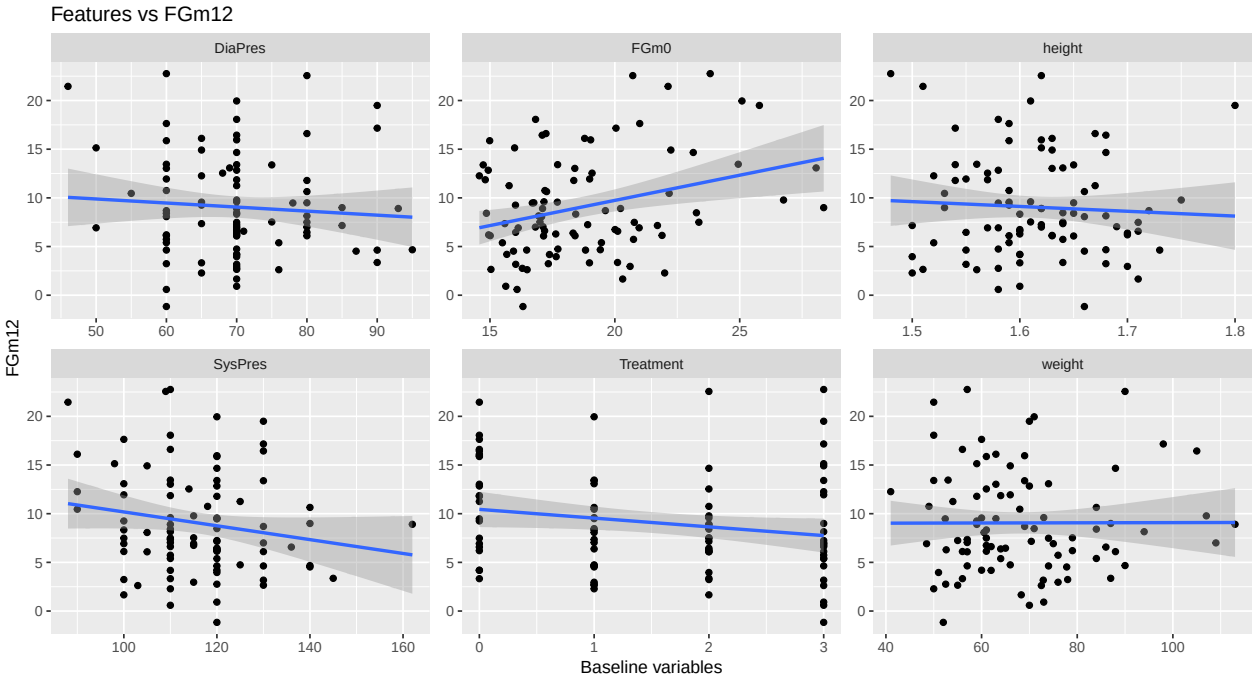
GAM fits for hirsutism data

Marcel Porta Valles, Javier Ferrando, Joan Prat

06/01/2020

We show a scatterplot of every variable measured at the beginning of the clinical trial against FGm12 (target variable) and a linear regression to show the tendency.



We can see that between FGm12 and FGm0 there's an apparent linear relationship while other features doesn't seem to have a clear linear correlation with our target variable.

Multiple Linear Regression

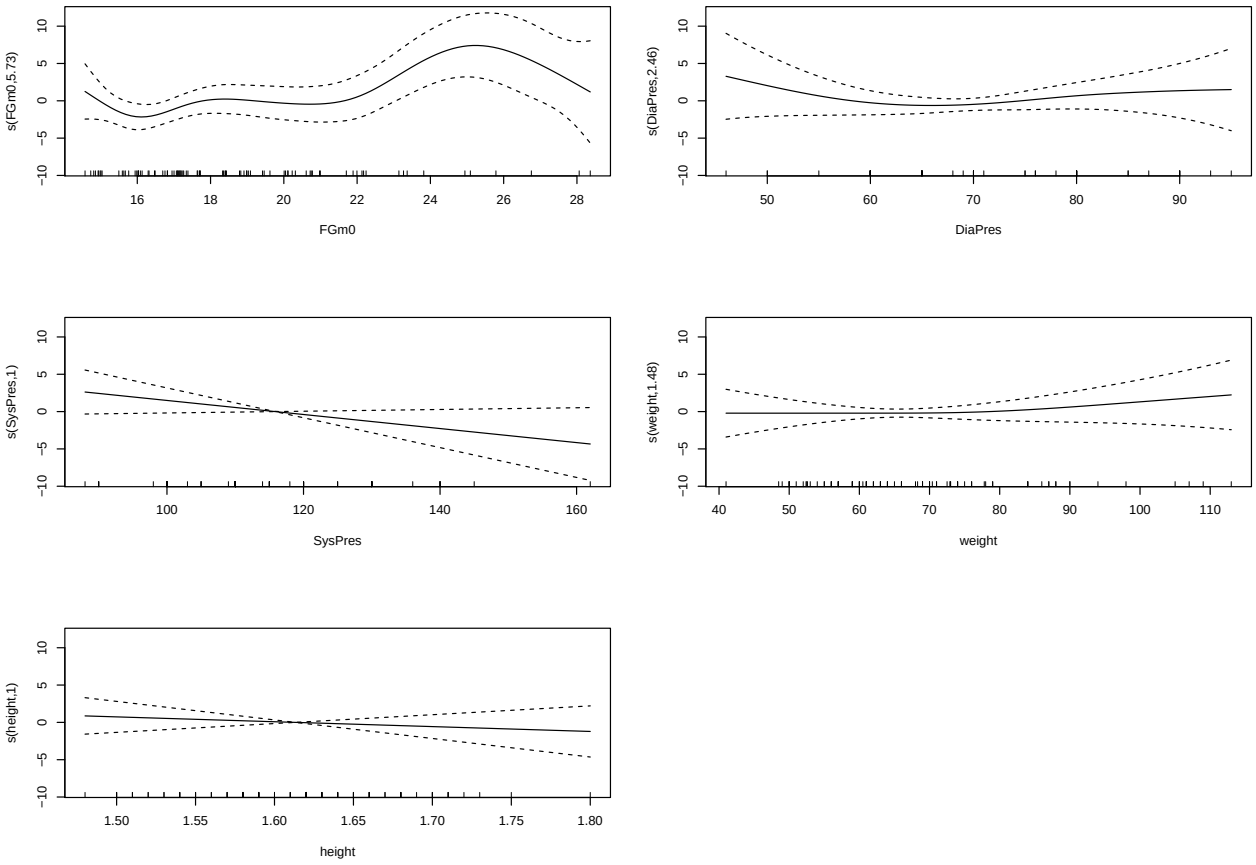
Firstly, we start with a simple multiple linear regression with every 'baseline' variable $y = \alpha + \beta_1 \cdot FGm0 + \beta_1 \cdot Treatment + \beta_2 \cdot DiaPres + \beta_3 \cdot SysPres + \beta_4 \cdot weight + \beta_5 \cdot height$ and observe that the p-values of the t-statistic for the coefficients of variables DiaPres, SysPres, weight and height lay above the 0.005 threshold. So, null hypothesis H_0 : There is no linear relationship between the prioir metnioned predictors and $FGm12$ can't be rejected.

	p.pv
(Intercept)	0.0801277
FGm0	0.0024259
Treatment	0.0151571
DiaPres	0.8503345
SysPres	0.1071206
weight	0.3644611
height	0.2010839

R-sq.(adj): 0.1343202

Generalized Additive Model using splines

$y = \alpha + s(FGm0) + s(DiaPres) + s(SysPres) + s(weight) + s(height)$



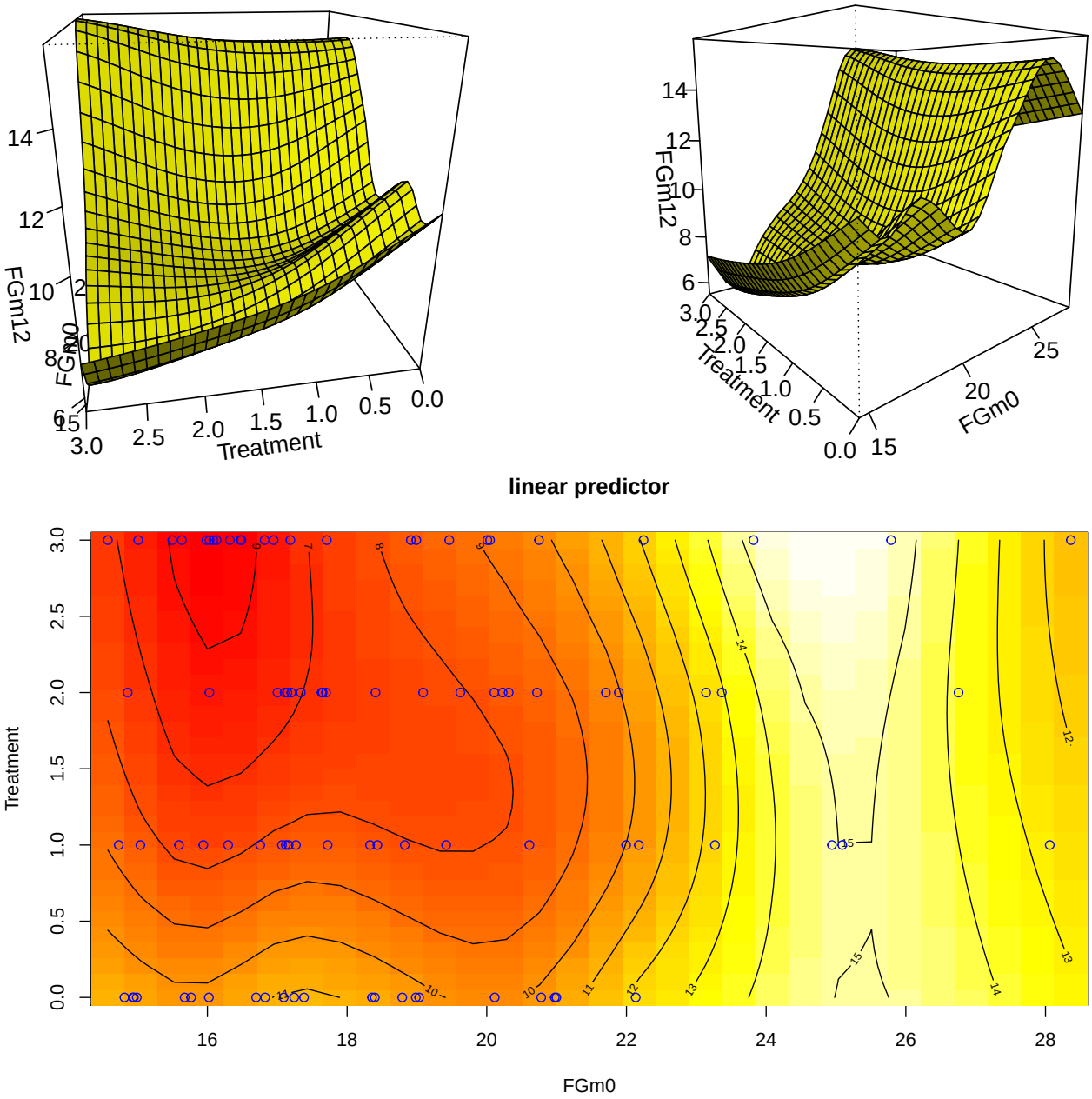
smooth_terms	s.pv
s(FGm0)	0.01853665
s(DiaPres)	0.40501150
s(SysPres)	0.07898182
s(weight)	0.58191465
s(height)	0.47998198

As it can be observed in the plots, spline function $s()$ finds as the best option best almost constant value functions, taking a look a the p-values, there is no clear evidence that a non-linear term is required for the 'baseline' variables except for *FGm0*.

R-sq.(adj): 0.1837405

Non-parametric bivariate regression using splines (thin plate)

$y = \alpha + s(FGm0, Treatment)$



The smoothing splines model with $FGm0$ and $Treatment$ shows that depending on the initial hirsutism value, a treatment might be better than the others. For example, for low values of $FGm0$, treatment 3 shows better results after 12 months ($FGm12$) than treatment 1 or 0. However, for higher values of $FGm0$, Treatment doesn't seem to be very decisive.

R-sq.(adj): 0.227637

Semiparametric model

$y = \alpha + s(FGm0, Treatment) + height + weight + SysPres$

After testing several combinations of non-parametric bivariate regression $s(FGm0, Treatment)$ together with linear combinations of 'baseline' variables, we can't find a significant improvement.

R-sq.(adj): 0.2602553

Model selection with ANOVA

Multiple Linear Regression (MLR) vs Generalized Additive Model using splines (GAMs)

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
84.0000	1985.265	NA	NA	NA	NA
76.2726	1745.733	7.7274	239.5324	1.390979	0.2160287

No evidence to reject H_0 :, so we accept MLR is a better model.

Generalized Additive Model using splines (GAMs) vs Bivariate regression using splines (Bis)

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
76.27260	1745.733	NA	NA	NA	NA
76.19513	1685.897	0.0774745	59.8355	36.62661	0.0049016

Evidence to reject H_0 :, so we accept Bis is a better model.

Bivariate regression using splines (Bis) vs Semiparametric model (SP)

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
76.19513	1685.897	NA	NA	NA	NA
73.84870	1564.018	2.346426	121.8789	2.571923	0.0744343

No evidence to reject H_0 :, so we accept Bis is a better model.

Multiple Linear Regression (MLR) vs Bivariate regression using splines (Bis)

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
84.00000	1985.265	NA	NA	NA	NA
76.19513	1685.897	7.804874	299.3679	1.819013	0.0880561

No evidence to reject H_0 :, so we accept MLR is a better model.

After comparing different models by pairs, we get that Multiple Linear Regression (MLR) looks the best model dispate it achieves lower R-sq.(adj) than the proposed Bivariate regression with splines.