GAM Fits for Hirsutism Data

Azzarito Domenico, Daniel Reverter, Alexis Vendrix

This document performs Generalized Additive Model (GAM) analysis on hirsutism clinical trial data. The goal is to model the Ferriman-Gallwey score at 12 months (`FGm12`) as a function of baseline measurements and treatment levels.

1. Data Loading and Cleaning

We begin by loading the hirsutism dataset and preparing it for analysis. This includes converting the treatment variable to a factor and handling missing or erroneous values.

```r
library(mgcv)
```

```r
# Load the dataset.
hirs <- read.table("hirsutism.dat", header = TRUE, sep = "\t", fill = TRUE)

# Convert Treatment to factor.
hirs$Treatment <- as.factor(hirs$Treatment)

# Correct erroneous negative FGm12 values.
neg_idx <- which(hirs$FGm12 < 0)
hirs$FGm12[neg_idx] <- 0

# Remove observations with missing values.
hirs <- na.omit(hirs)
```

Table 1: Treatment Group Counts

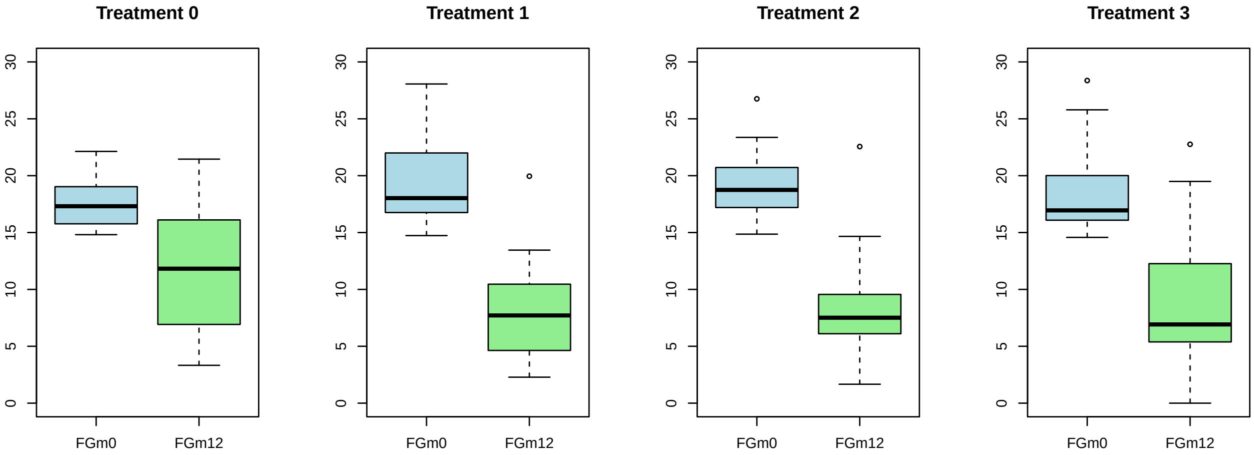
Value	Frequency
0	22
1	22
2	22
3	25

Table 2: Summary Statistics of Key Variables

	FGm0	FGm12	SysPres	DiaPres	weight	height
	Min. :14.57	Min. : 0.000	Min. : 88.0	Min. :46.00	Min. : 41.00	Min. :1.480
	1st Qu.:16.40	1st Qu.: 5.566	1st Qu.:110.0	1st Qu.:65.00	1st Qu.: 57.00	1st Qu.:1.580
	Median :17.70	Median : 8.069	Median :115.0	Median :70.00	Median : 64.00	Median :1.610
	Mean :18.67	Mean : 9.066	Mean :115.9	Mean :70.04	Mean : 68.06	Mean :1.613
	3rd Qu.:20.27	3rd Qu.:12.402	3rd Qu.:120.0	3rd Qu.:75.00	3rd Qu.: 74.50	3rd Qu.:1.650
	Max. :28.36	Max. :22.759	Max. :162.0	Max. :95.00	Max. :113.00	Max. :1.800

2. Exploratory Data Analysis

Here, we explore the relationships between predictors and the response variable through visualizations and correlation analysis.



```r
# Compute correlations between numeric predictors and FGm12.
numeric_vars <- hirs[, c("FGm0", "SysPres", "DiaPres",
                         "weight", "height", "FGm12")]
cor_with_target <- cor(numeric_vars, method = "pearson")[, "FGm12"]
```

Table 3: Pearson Correlations with FGm12

Variable	Correlation
FGm0	0.308
SysPres	-0.177
DiaPres	-0.079
weight	0.000
height	-0.057

The correlations are relatively low, suggesting that linear relationships alone may not capture the underlying patterns. This motivates the use of GAMs with smooth terms.

3. Model Fitting

We fit several GAM models of increasing complexity, starting from a simple linear model and progressively adding smooth terms and interactions.

```r
set.seed(123)
```

```r
# Model 0: Linear model (baseline).
gam0 <- gam(FGm12 ~ FGm0 + SysPres + DiaPres + weight + height + Treatment,
            data = hirs)
```

```r
# Model 1: Full additive model with smooth terms.
gam1 <- gam(FGm12 ~ s(FGm0) + s(SysPres) + s(DiaPres) + s(weight) + s(height)
            + Treatment, data = hirs)

# Model 2: Treatment-specific smooths for FGm0.
gam2 <- gam(FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + s(DiaPres) + s(weight)
            + s(height) + Treatment, data = hirs)
```

```r
# Model 3: Reduced model (remove non-significant DiaPres).
gam3 <- gam(FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + s(weight) + s(height)
            + Treatment, data = hirs)

# Model 4: Further reduction (remove height).
gam4 <- gam(FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + s(weight) + Treatment,
            data = hirs)

# Model 5: Minimal smooth model.
gam5 <- gam(FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + Treatment,
            data = hirs)

# Model 6: Only FGm0 smooth by Treatment.
gam6 <- gam(FGm12 ~ s(FGm0, by = Treatment) + Treatment,
            data = hirs)
```

```r
# Model 7: Tensor product for DiaPres and weight.
gam7 <- gam(FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) +
              te(DiaPres, weight) + Treatment, data = hirs)

# Model 8: Two tensor products (best candidate).
gam8 <- gam(FGm12 ~ s(FGm0, by = Treatment) + te(DiaPres, weight) +
              te(SysPres, height) + Treatment, data = hirs)
```

4. Model Selection

We compare models using GCV scores and ANOVA to identify the best-fitting model.

```r
# Extract GCV scores for comparison.
gcv_scores <- c(
  Linear = gam0$gcv.ubre,
  GAM1 = gam1$gcv.ubre,
  GAM2 = gam2$gcv.ubre,
  GAM3 = gam3$gcv.ubre,
  GAM4 = gam4$gcv.ubre,
  GAM5 = gam5$gcv.ubre,
  GAM6 = gam6$gcv.ubre,
  GAM7 = gam7$gcv.ubre,
  GAM8 = gam8$gcv.ubre
)
```

Table 4: GCV Scores Across Models (Lower is Better)

Model	GCV
Linear	24.950
GAM1	23.146
GAM2	22.563
GAM3	22.562
GAM4	22.086
GAM5	22.815
GAM6	23.021
GAM7	21.608
GAM8	20.922

```r
# ANOVA comparison of nested models.
anova(gam5, gam7, gam8, gam4, gam3, gam2, gam0, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + Treatment
## Model 2: FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + te(DiaPres, weight) +
##     Treatment
## Model 3: FGm12 ~ s(FGm0, by = Treatment) + te(DiaPres, weight) + te(SysPres,
##     height) + Treatment
## Model 4: FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + s(weight) + Treatment
## Model 5: FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + s(weight) + s(height) +
##     Treatment
## Model 6: FGm12 ~ s(FGm0, by = Treatment) + s(SysPres) + s(DiaPres) + s(weight) +
##     s(height) + Treatment
## Model 7: FGm12 ~ FGm0 + SysPres + DiaPres + weight + height + Treatment
##   Resid. Df Resid. Dev       Df Deviance      F    Pr(>F)
## 1    71.641    1378.00
## 2    57.537     904.26  14.1039   473.74 2.7106 0.0049831 **
## 3    48.856     667.87   8.6818   236.39 2.1973 0.0399574 *
## 4    66.245    1175.31 -17.3892  -507.44 2.3549 0.0097072 **
## 5    67.934    1245.75  -1.6895   -70.44 3.3647 0.0503650 .
## 6    63.890    1122.11   4.0448   123.64 2.4669 0.0564517 .
## 7    82.000    1843.55 -18.1105  -721.44 3.2147 0.0005985 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table reveals significant improvements when moving from simpler to more complex models. Notably, the transition from GAM5 to GAM7 (adding `te(DiaPres, weight)`) and from GAM7 to GAM8 (adding `te(SysPres, height)`) both show significant F-statistics (p < 0. 05). The comparison also confirms that the linear model (gam0) is significantly outperformed by the GAM alternatives (p < 0.001).

Based on the lowest GCV score (20.922) and ANOVA results, **Model 8** (`gam8`) is selected as the best model. It achieves 72.6% deviance explained through treatment-specific smooths for `FGm0` and tensor product interactions.

5. Final Model Examination

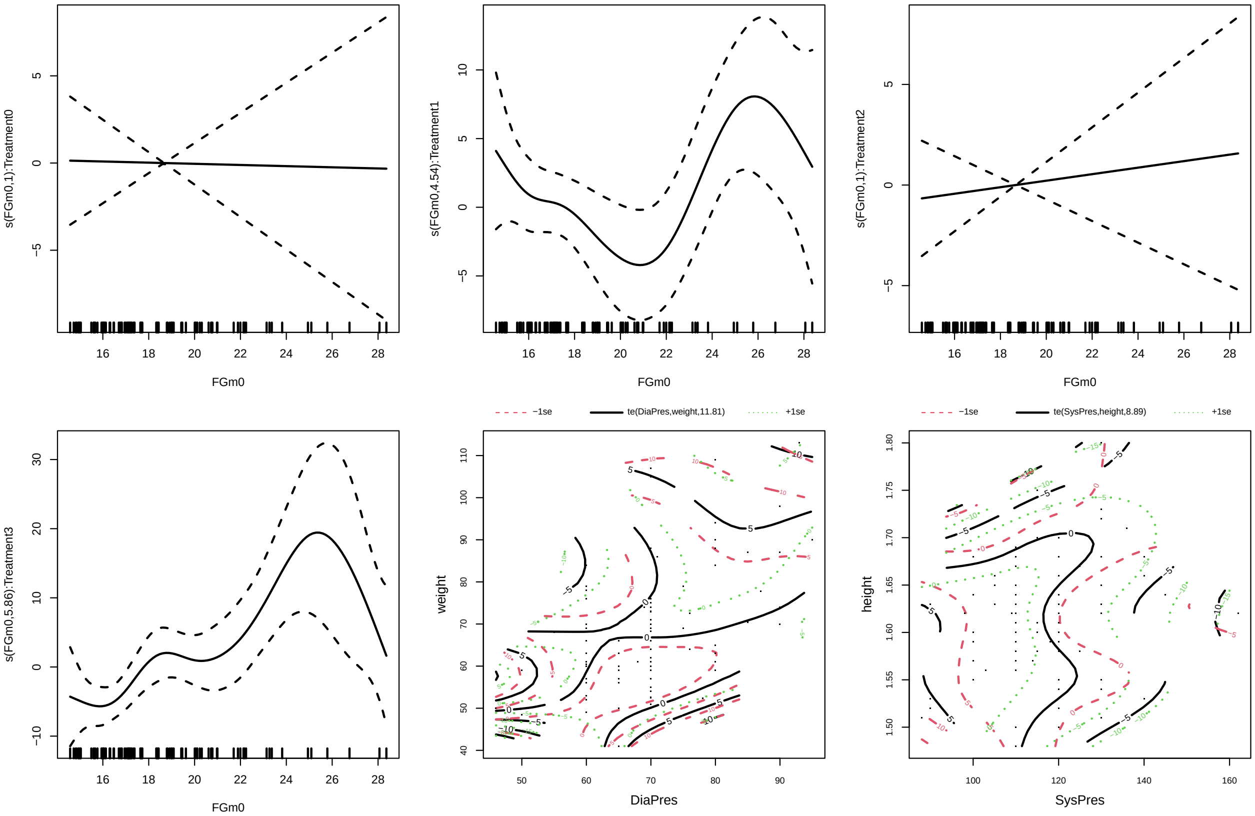
We examine the selected model in detail using summary statistics, smooth plots, and diagnostic checks.

```r
summary(gam8)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## FGm12 ~ s(FGm0, by = Treatment) + te(DiaPres, weight) + te(SysPres,
##     height) + Treatment
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.1839     0.9629  11.615 2.71e-16 ***
## Treatment1   -3.4750     1.3632  -2.549   0.0137 *
## Treatment2   -1.8591     1.3216  -1.407   0.1653
## Treatment3   -2.8704     1.3489  -2.128   0.0379 *
## ---
## 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):Treatment0   1.000  1.000 0.006 0.941025
## s(FGm0):Treatment1   4.540  5.395 2.495 0.036942 *
## s(FGm0):Treatment2   1.000  1.000 0.215 0.644806
## s(FGm0):Treatment3   5.864  6.824 4.834 0.000398 ***
## te(DiaPres,weight) 11.814 13.226 2.293 0.016420 *
## te(SysPres,height)  8.885 10.700 1.977 0.056693 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.542   Deviance explained = 72.6%
## GCV = 20.922  Scale est. = 12.392    n = 91
```

The model summary reveals:

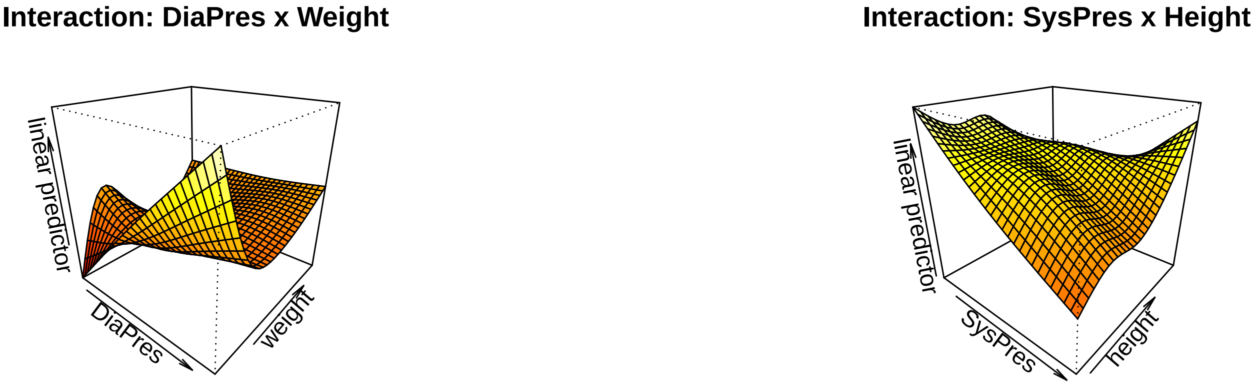
- **Parametric coefficients**: Treatment1 (p = 0.014) and Treatment3 (p = 0.038) show significant reductions in FGm12 scores relative to the control group. Treatment2 does not reach statistical significance (p = 0.165).The `te(SysPres,height)` is borderline significant but we will consider it because we show afterward that there is an improvement from Gam7 to Gam8.
- **Smooth terms**: `s(FGm0):Treatment1` and `s(FGm0):Treatment3` are significant, indicating nonlinear baseline effects for these treatment groups. The tensor product `te(DiaPres, weight)` is significant (p = 0.016), while `te(SysPres, height)` is marginally significant (p = 0.057).

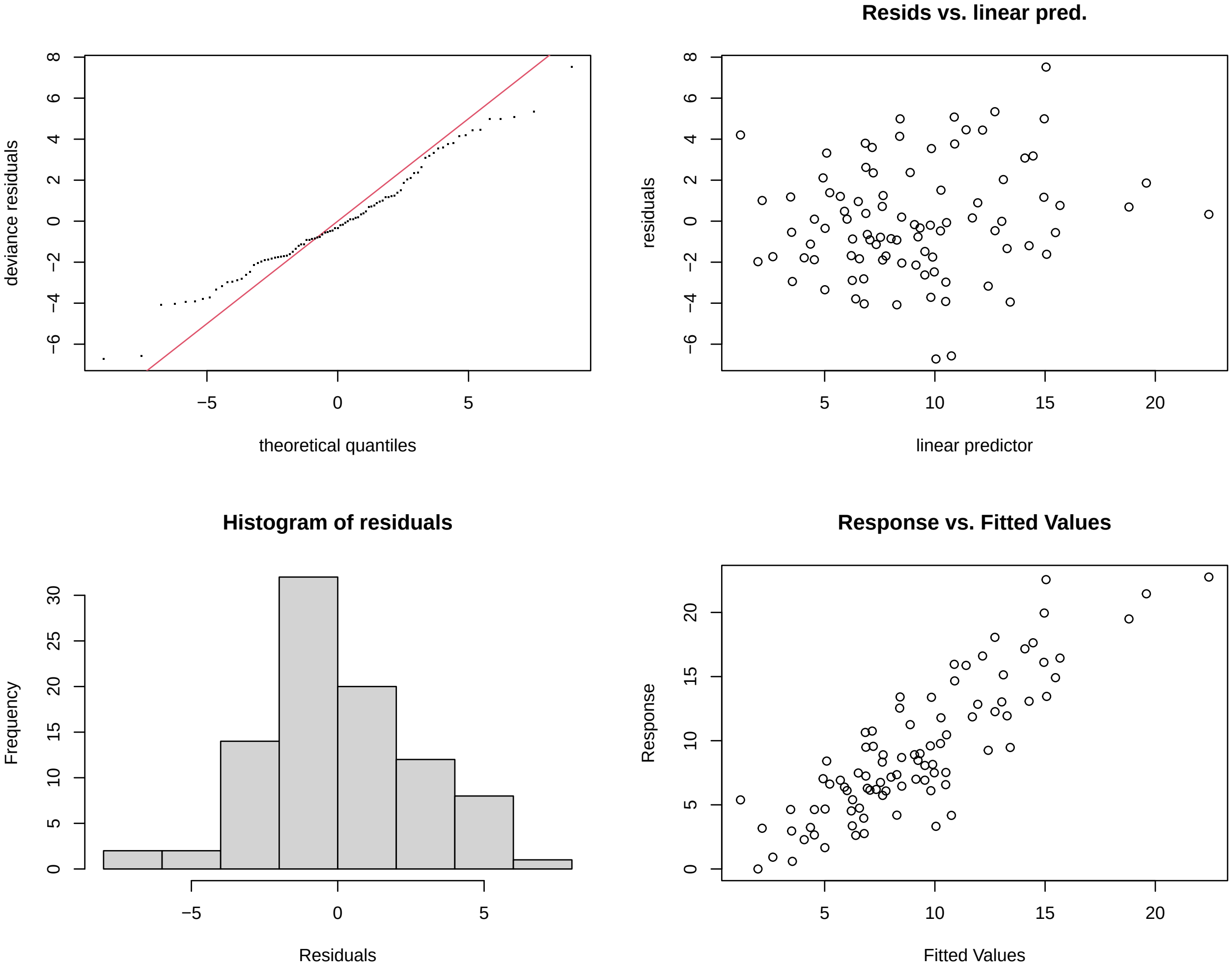


The smooth plots reveal:

- **Treatment 0 and 2**: Flat effects of `FGm0` (edf ≈ 1), indicating essentially linear or no relationship between baseline hirsutism and final outcome for these groups.

- **Treatment 1**: A nonlinear pattern (edf = 4.54) with higher `FGm12` values at elevated baseline levels.
- **Treatment 3**: The most complex relationship (edf = 5.86) showing strong nonlinear dependence on baseline severity.
- **Tensor products**: Capture complex interactions, with `te(DiaPres, weight)` using 11.8 effective degrees of freedom.





```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 18 iterations.
## The RMS GCV score gradient at convergence was 1.072644e-06 .
## The Hessian was positive definite.
## Model rank =  88 / 88
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                          k'   edf k-index p-value
## s(FGm0):Treatment0  9.00  1.00    1.05   0.665
## s(FGm0):Treatment1  9.00  4.54    1.05   0.630
## s(FGm0):Treatment2  9.00  1.00    1.05   0.690
## s(FGm0):Treatment3  9.00  5.86    1.05   0.660
## te(DiaPres,weight) 24.00 11.81    1.07   0.825
## te(SysPres,height) 24.00  8.89    0.85   0.025 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The diagnostic plots indicate:

- **Q-Q plot**: Residuals follow theoretical quantiles reasonably well, with minor deviations at the tails.
- **Residuals vs. fitted**: No systematic patterns, suggesting homoscedasticity.
- **Histogram**: Approximately normal distribution of residuals.
- **Response vs. fitted**: Positive correlation between observed and predicted values.

The basis dimension check shows adequate k values for most smooth terms (k-index $>$ 1). However, `te(SysPres, height)` displays a k-index of 0.86 with p = 0.03, suggesting potential under-smoothing and over-fitting given the sample size of 91 observations. Since the effective degrees of freedom (edf = 8.89) is well below the basis dimension (k' = 24), this is likely a minor concern and no corrective action is required.

6. Concluding Remarks

The Generalized Additive Model analysis demonstrates that smooth terms and tensor product interactions substantially improve model fit compared to a simple linear model. The final model (`gam8`) achieves the lowest GCV score (20.922) and explains 72.6% of the variability in `FGm12` by:

- Modeling treatment-specific nonlinear effects of baseline hirsutism levels (`FGm0`).
- Capturing interactions between physiological variables (`DiaPres × weight` and `SysPres × height`) through tensor product smooths.

Key findings from the clinical perspective:

- **Treatment efficacy varies**: Treatments 1 and 3 show statistically significant reductions in Ferriman-Gallwey scores relative to the control group, while Treatment 2 does not reach significance.
- **Baseline severity matters differently by treatment**: For Treatments 0 and 2, baseline hirsutism has minimal impact on final outcomes. For Treatments 1 and 3, participants with higher baseline FG values show stronger nonlinear responses, suggesting these treatments may be more effective for patients with moderate baseline severity.
- **Physiological interactions**: The significant interaction between diastolic blood pressure and weight suggests that body composition and cardiovascular factors may influence treatment response, warranting further clinical investigation.