University of Groningen

Mathematics

Generalized Additive Models and its Application

Abstract

Generalized additive modeling is a semi-parametric method of regression modeling. A generalized additive model is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables. The model allows for rather flexible specification of the dependence of the response on the covariates. The example about a common shrub in the Bryce Canyon National Park will demonstrate the potential benefits of using generalized additive models in comparison with generalized linear models. Our final goal was to apply generalized additive models to medical statistics and find new associations. Unfortunately, no new significant association is found.

Author:
Teyler Kroon

First Supervisor:
Dr. W. P. Krijnen

Second Supervisor:
Dr. M. A. Grzegorczyk

March 28, 2017

Contents

1 Introduction **2**

2 Generalized Additive Models **3**
2.1 Univariate Smooth Functions 3
2.1.1 Controlling the degree of smoothing with penalized regression splines 4
2.1.2 Choosing the smoothing parameter, λ: cross validation . . 5
2.2 Additive Models 7
2.2.1 Penalized regression spline representation of an additive model 7
2.2.2 Fitting additive models by penalized least squares 8
2.3 Generalized Additive Models 8

3 Modeling Environment Relations **10**
3.1 Modeling Environment Relations with GLMs 10
3.2 Modeling Environment Relations with GAMs 12
3.3 GLMs versus GAMs 15

4 GAM Application on a Medical Article **17**
4.1 Functional Somatic Symptoms Related to Perceived Stress 17
4.2 Functional Somatic Symptoms Related to Heart Rate Variability and Pre-Ejection Period 22

5 Epilogue **26**

Appendix A **28**

Appendix B **29**

1 Introduction

Regression analysis is a branch of statistics that examines the relationship between various variables. In this paper, we investigate generalized additive models (GAMs), a semi-parametric method of regression modeling. Sometimes, parametric models do not quite fit the data. In such cases the smoothing technique of generalised additive models could be the solution. The design and development of smoothers is a very active area of research in statistics [4].

This thesis describes the theory of GAMs and contains an example wherein we model environment relations. The example about a common shrub in the Bryce Canyon National Park will demonstrate the potential benefits of using GAMs in comparison with generalized linear models. Linear modeling is applied frequently in all sort of branches and we presume that some of these regression analysis will have a better fit with GAMs. Therefore, the literature will be secluded by an application of GAMs on a medical research wherein the data is analyzed by linear regression. The publication examines the associations between stressors and functional somatic symptons such as headaches and overtiredness. A total of 36 linear analyzes are performed which we will examine by GAMs in R.

2 Generalized Additive Models

In statistics, a generalized additive model [5] is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables. The model relates a univariate response variable, Y_i, to some predictor values x_k. In general the model has a structure something like

$$g(\mu_i) = \mathbf{X}_i^* \theta + f_1(x_{1_i}) + f_2(x_{2_i}) + f_3(x_{3_i}, x_{4_i}) + ... \tag{1}$$

where

$$\mu_i \equiv \mathbb{E}(Y_i) \quad \text{and} \quad Y_i \sim \text{some exponential family distribution}$$

Our g is a known smooth monotonic function, $\mathbf{X}_i^*$ is a row of the model matrix for any strictly parametric model components, θ is the corresponding parameter vector, and the f_j are smooth functions of the predictor variables, x_k. The model allows for rather flexible specification of the dependence of the response on the covariates, but by specifying the model only in terms of 'smooth funcions', rather than detailed parametric relationships, it is possible to avoid the sort of cumbersome and unwieldy models. This flexibility and convenience comes at the cost of two new theoretical problems. It is necessary both to represent the smooth functions in some way and to choose how smooth they should be. [8]

2.1 Univariate Smooth Functions

The representation of smooth functions is best introduced by considering a model containing one smooth function of one covariate:

$$y_i = f(x_i) + \epsilon_i \tag{2}$$

where y_i is a response variable, x_i a covariate, f a smooth function and the ϵ_i are i.i.d. $N(0, \sigma^2)$ random variables. To further simplify matters, suppose in chapter 2 that the covariates lie in the interval $[0, 1]$.

To estimate f we want our smooth funtion to be a linear model. Therefore we will choose a basis for f which defines the space of functions of which f is an element. When we choose a basis we also have the basis functions, which we can treat as completely known. If $b_i(x)$ is the i^{th} such basis function, then f is assumed to have a representation

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i \tag{3}$$

for some values of the unknown parameters, β_i. If we substitute this respresentation of f into our smooth function it clearly yields our desired linear model.

A univariate function can be represented using a cubic spline. A cubic spline is a curve, made up of sections of cubic polynomial, joined together so that they are continuous in value as well as first and second derivatives. The points at which the sections join are known as the knots of the spline. The knots occur wherever there is a datum. Let the knot locations be denoted by $x_i^* : i = 1, ..., q - 2$.

Given knot locations, there are many alternative, but equivalent, ways of writing down a basis for cubic splines. Our simple basis results from a very general

approach to splines by Wahba [6] and Gu [1], although it looks like a very complicated form. Our basis functions for this basis in the general approach are chosen as: $b_1(x) = 1, b_2(x) = x$ and $b_{i+2} = R(x, x_i^*)$ for $i = 1...q-2$ where

$$R(x,z) = \Big[(z-1/2)^2 - 1/12\Big]\Big[(x-1/2)^2 - 1/12\Big]/4 - \Big[(|x-z|-1/2)^4 - 1/2(|x-z|-1/2)^2 + 7/240\Big]/24 \quad (4)$$

Using this cubic spline basis for f means that (2) becomes a linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where the i^{th} row of the model matrix is:

$$\mathbf{X}_i = [1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), ..., R(x_i, x_{q-2}^*)]$$

Hence the model can be estimated by least squares.

2.1.1 Controlling the degree of smoothing with penalized regression splines

A way to controlling smoothness by altering the basis dimension, is to keep the basis dimension fixed, at a size a little larger than it is believed could reasonably be necessary, but to control the model's smoothness by adding "wiggliness" penalty to the least squares fitting objective. For example, rather than fitting the model by minimizing,

$$\|\mathbf{y} - \mathbf{X}\beta\|^2, \quad (5)$$

it could be fit by minimizing,

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \int_0^1 [f''(x)]^2 dx, \quad (6)$$

where the integrated square of second derivative penalizes models that are too "wiggly". The trade off between model fit and model smoothness is controlled by the smoothing parameter, λ. $\lambda \to \infty$ renders curvature impossible, thereby returning us to ordinary linear regression which is described by one parameter. $\lambda = 0$ imposes no restrictions and results in an un-penalized regression spline estimate such that the effective number of parameters is greater than one.

Because f is linear in the parameters, β_i, the penalty can always be written as a quadratic form in β. Differentiating the basis expansion of f in (3), we have $f''(x) = \beta^T \mathbf{d}(\mathbf{x})$ where $d_i(x) = b_i''(x)$. Using the fact that a scalar is its own transpose we then have that:

$$\int_0^1 [f''(x)]^2 dx = \int_0^1 [\beta^{\mathbf{T}} \mathbf{d}(\mathbf{x}) \mathbf{d}(\mathbf{x})^{\mathbf{T}} \beta] dx = \beta^{\mathbf{T}} \mathbf{S} \beta \quad (7)$$

where $\mathbf{S}$ is a matrix of known coefficients. It is now that the somewhat complicated form of the spline basis, used here, proves its worth, for it turns out that $S_{i+2,j+2} = R(x_i^*, x_j^*)$ for $i, j = 1, ..., q-2$ and x_k^* the knot locations. While the first two rows and columns of $\mathbf{S}$, which are the multiplied square of the second derivatives of basis functions b_1 and b_2, are 0. A property of our matrix $\mathbf{S}$ is, that it is symmetric since $\mathbf{S} = \mathbf{d}(\mathbf{x})\mathbf{d}(\mathbf{x})^{\mathbf{T}}$ with $d_i(x) = b_i''(x)$. Since we defined our basis functions with the properties that the variables are in the non-negative

domain our matrix is positive semi-definite. By definition we have that a function is convex if and only if $\mathbf{S}$ is symmetric and positive semi-definite. Therefore our function is convex.

By (7) and shown in (6), the penalized regression spline fitting problem now is to minimize

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\beta^T\mathbf{S}\beta \tag{8}$$

w.r.t. β. Similarly, for practical computation, if $\mathbf{B}$ is any square root of the matrix $\mathbf{S}$ such that $\mathbf{B^T B} = \mathbf{S}$, note that

$$\left|\left| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{B} \end{bmatrix} \beta \right|\right|^2 = ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda\beta^T\mathbf{S}\beta \tag{9}$$

The problem of estimating the degree of smoothness for the model is now the problem of estimating the smoothing parameter λ. But before addressing λ estimation, consider β estimation, given λ,

$$\begin{aligned} S_p &= \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\beta^{\mathbf{T}}\mathbf{S}\beta \\ &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^{\mathbf{T}}\mathbf{S}\beta \\ &= \mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \hat{\beta}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})\beta \end{aligned}$$

Differentiating S_p w.r.t. β and setting to zero results in the system of equations

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})\hat{\beta} = \mathbf{X}^T\mathbf{y},$$

which yields the result of the penalized least squares estimator of β

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T\mathbf{y} \tag{10}$$

2.1.2 Choosing the smoothing parameter, λ: cross validation

If the smoothing paramater is too high then the data will be over smoothed, and the data will be under smoothed if λ is too low. In both cases this will mean that the spline estimate $\hat{f}$ deviates from the true function f. Ideally, we can choose λ so that $\hat{f}$ is as close as possible to f. A suitable criterion might be to choose λ to minimize

$$M = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2, \tag{11}$$

where the notation $\hat{f}_i \equiv \hat{f}(x_i)$ and $f_i \equiv f(x_i)$ have been adopted for conciseness. Since f is unknown, M cannot be used directly, but it is possible to derive an estimate of $\mathbb{E}(M) + \sigma^2$, which is the expected squared error in predicting a new variable. Let $\hat{f}^{[-i]}$ denotes the prediction of $\mathbb{E}(y_i)$ obtained from the model fitted to all data except y_i, and define the ordinary cross validation score

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2 \tag{12}$$

This score results from leaving out each datum in turn, fitting the model to the remaining data and calculating the squared difference between the missing datum and its predicted value: these squared differences are then averaged over all the data. Substituting $y_i = f_i + \epsilon_i$,

$$\begin{aligned}\mathcal{V}_o &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2 - (\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2.\end{aligned}$$

Since $\mathbb{E}(\epsilon_i) = 0$, and ϵ_i and $\hat{f}_i^{[-i]}$ are independent, the second term in the summation vanishes if expectations are taken:

$$\mathbb{E}(\mathcal{V}_o) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\right) + \sigma^2. \tag{13}$$

Now, $\hat{f}_i^{[-i]} \approx \hat{f}$ with equality in the large sample limit, so $\mathbb{E}(\mathcal{V}_o) \approx \mathbb{E}(M) + \sigma^2$ also with equality in the large sample limit. Hence choosing λ in order to minimize $\mathcal{V}_o$ is a reasonable approach if the ideal would be to minimize M. Choosing λ to minimize $\mathcal{V}_o$ is known as ordinary cross validation(OCV).

Fortunately, calculating $\mathcal{V}_o$ by performing n models first, to obtain the n terms $\hat{f}_i^{[-i]}$, is unnecessary. To see this, first consider the penalized least squares objective which in principle has to be minimized to find the i^{th} term in the OCV score:

$$\sum_{\substack{j=1\\ j\neq i}}^{n}(y_j - \hat{f}_j^{[-i]}) + C$$

with C the penalties. Clearly, adding zero to this objective will leave the estimates that minimize it completely unchanged. So we can add the term $(\hat{f}_i^{[-i]} - \hat{f}_i^{[-i]})^2$ to obtain

$$\sum_{j=1}^{n}(y_j^* - \hat{f}_j^{[-i]})^2 + C \tag{14}$$

Where $\mathbf{y}^* = \mathbf{y} - \bar{\mathbf{y}}^{[\mathbf{i}]} + \bar{\mathbf{f}}^{[\mathbf{i}]}$: $\bar{\mathbf{y}}^{[\mathbf{i}]}$ and $\bar{\mathbf{f}}^{[\mathbf{i}]}$ are vectors of zeroes except for their i^{th} elements which are y_i and $\hat{f}_i^{[-i]}$, respectively.

Fitting, by minimizing (14), obviously results in i^{th} prediction $\hat{f}_i^{[-i]}$, and also in an influence matrix $\mathbf{A}$, which is just the influence matrix for the model fitted to all the data. Note that $\hat{y} = \mathbf{A}y = \mathbf{X}\hat{\beta}$ and from (10) we can define our influence matrix as $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{S})^{-1}\mathbf{X}^T$. So considering the i^{th} prediction we have that:

$$\hat{f}_i^{[-i]} = \mathbf{A_i}\mathbf{y}^* - A_{ii}y_i + A_{ii}\hat{f}_i^{[-i]} = \hat{f}_i - A_{ii}y_i + A_{ii}\hat{f}_i^{[-i]},$$

Where $\hat{f}_i$ is from the fit to the full $\mathbf{y}$. Substraction of y_i, from both sides, and a little rearrangement then yields

$$y_i - \hat{f}_i^{[-i]} = (y_i - \hat{f}_i)/(1 - A_{ii})$$

So that the OCV score becomes

$$\mathcal{V}_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2/(1 - A_{ii})^2 \tag{15}$$

Which can clearly be calculated from a single fit of the original model. In practice the weights, $1-A_{ii}$, are often replaced by the mean weight, $tr(\mathbf{I} - \mathbf{A})/n$, in order to arrive at the generalized cross validation (GCV) score

$$\mathcal{V}_g = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{[tr(\mathbf{I} - \mathbf{A})]^2} \tag{16}$$

2.2 Additive Models

Now suppose that two predictable variables, x and z, are available for a response variable, y, and that a simple additive model structure

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i \tag{17}$$

is appropriate. The f_j are smooth functions, and the ϵ_i are i.i.d. $N(0, \sigma^2)$ random variables.

There are two points to note about this model. Firstly, the assumption of additive effects is a fairly strong one: $f_1(x) + f_2(z)$ is a quite restrictive special case of the general smooth function of two variables $f(x, z)$. Secondly, the fact that the model now contains more than one function introduces an identifiability problem: f_1 and f_2 are each only estimable to within an additive constant. To see this, note that any constant could be simultaneously added to f_1 and subtracted from f_2, without changing the model predictions. Hence identifiability constraints have to be imposed on the model before fitting.

Provided the identifiability issue is addressed, the additive model can be represented using penalized regression splines, estimated by penalized least squares and the degree of smoothing estimated by cross validation, in the same way as the simple univariate model.

2.2.1 Penalized regression spline representation of an additive model

Each smooth function of our simple additive model structure can be represented using a penalized regression spline basis. Using the spline basis from (4) we get

$$f_1(x) = \delta_1 + x\delta_2 + \sum_{j=1}^{q_1-2} R(x, x_j^*)\delta_{j+2} \tag{18}$$

and

$$f_2(z) = \gamma_1 + z\gamma_2 + \sum_{j=1}^{q_2-2} R(z, z_j^*)\gamma_{j+2} \tag{19}$$

where δ_j and γ_j are the unknown parameters for f_1 and f_2 respectively. q_1 and q_2 are the number of unknown parameters for f_1 and f_2, while x_j^* and z_j^* are

the knot locations for the two functions.

The identifiability problem with the additive model means that δ_1 and γ_1 are confounded. The simplest way to deal with this is to constrain one of them to zero, say $\gamma_1 = 0$. Having done this, it is easy to see that the additive model can be written in the linear model form, $y = X\beta + \epsilon$, where the i^{th} row of the model matrix is now

$$X_i = [1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), ..., R(x_i, x_{q_1-2}^*), z_i, R(z_i, z_1^*), ..., R(z_i, z_{q_2-2}^*)] \tag{20}$$

and the parameter vector is $\beta = [\delta_1, \delta_2, ..., \delta_{q_1}, \gamma_2, \gamma_3, ..., \gamma_{q_2}]^T$

Exactly as shown in (7) we can also measure the wiggliness of the functions by

$$\int_0^1 f_1''(x)^2 dx = \beta^T \mathbf{S}_1 \beta \quad \text{and} \quad \int_0^1 f_2''(x)^2 dx = \beta^T \mathbf{S}_2 \beta \tag{21}$$

where $\mathbf{S_1}$ and $\mathbf{S_2}$ are zero everywhere except for $\mathbf{S_1}_{i+2,j+2} = R(x_i^*, x_j^*)$ for $i, j = 1, ..., q_1 - 2$ and $\mathbf{S_2}_{i+q_1+1,j+q_1+1} = R(z_i^*, z_j^*)$ for $i, j = 1, ..., q_2 - 2$

It is, of course, perfectly possible to use any of a large number of alternative bases in place of the regression spline basis used here. Only the details are changed by doing this, not the general principle that, once a basis has been chosen, model matrices and penalty coefficient matrices can immidiately be obtained.

2.2.2 Fitting additive models by penalized least squares

The parameters β of the model (17) are obtained by minimization of the penalized least squares objective

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \beta^T \mathbf{S_1} \beta + \lambda_2 \beta^T \mathbf{S_2} \beta, \tag{22}$$

where the smoothing parameters λ_1 and λ_2 control the weight to be given to the objective of making f_1 and f_2 smooth, relative to the objective of closely fitting the response data. For the moment, assume that these smoothing paramaters are given. Defining $\mathbf{S} \equiv \lambda_1 \mathbf{S_1} + \lambda_2 \mathbf{S_2}$, the objective can be re-written as

$$||\mathbf{y} - \mathbf{X}\beta||^2 + \beta^T \mathbf{S} \beta = \left|\left| \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix} \beta \right|\right|^2 \tag{23}$$

Where $\mathbf{B}$ is any matrix square root such that $\mathbf{B^T B} = \mathbf{S}$. As in the single smooth case, the right hand expression is simply the un-penalized least squares objective for an augmented version of the model and corresponding response data: hence the model can be fitted by standard linear regression.

2.3 Generalized Additive Models

Generalized additive models follow from additive models, as generalized linear models follow from linear models. That is, the linear predictor now predicts some known smooth monotonic function of the expected value of the response, and the response may follow any exponential family distribution, or simply have

a known mean variance relationship, permitting the use of a quasi-likelihood approach. The resulting model has a general form as we have shown in (1).

Whereas the additive model was estimated by penalized least squares, the GAM will be fitted by penalized likelihood maximization: in practice this will be achieved by penalized iterative least squares, but there is no simple trick to produce an unpenalized GLM whose likelihood is equivalent to the penalized likelihood of the GAM that we wish to fit.

To fit the model we simply iterate the following penalized iteratively re-weighted least squares (P-IRLS) scheme to convergence.

1. Given current parameter estimates $\beta^{[k]}$, and corresponding estimated mean response vector $\mu^{[k]}$, calculate:

$$w_i \propto \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})} \quad \text{and} \quad z_i = g(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + \mathbf{X}_i\beta^{[k]}$$

where $\text{var}(Y_i) = V(\mu^{[k]})\phi$

2. Minimize

$$\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\beta)\|^2 + \lambda_1\beta^T\mathbf{S}_1\beta + \lambda_2\beta^T\mathbf{S}_2\beta$$

w.r.t. β to obtain $\beta^{[k+1]}$. $\mathbf{W}$ is a diagonal matrix such that $W_{ii} = w_i$

Step 2 can be replaced by the equivalent:

2a. Minimize

$$\left\| \begin{bmatrix} \sqrt{\mathbf{W}} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \left(\begin{bmatrix} \mathbf{z} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix} \beta \right) \right\|^2$$

w.r.t β to obtain $\beta^{[k+1]}$, where $\mathbf{B}$ is a matrix square root such that $\mathbf{B}^T\mathbf{B} = \lambda_1\mathbf{S}_1 + \lambda_2\mathbf{S}_2$

In the current case, the link function, g, is the *log*, so $g'(\mu_i) = \mu_i^{-1}$, while for the gamma distribution, $V(\mu_i) = \mu_i^2$. Hence, for the log-link, gamma errors model, we have:

$$w_i = 1 \quad \text{and} \quad z_i = \left(y_i - \mu_i^{[k]}\right)/\mu_i^{[k]} + \mathbf{X}_i\beta^{[k]}$$

3 Modeling Environment Relations

This chapter contains an example in which the potential advantage in using a GAM function is shown relative to a GLM function. Dave Roberts of the Department of Ecology of the Montana State University developed R labs for vegetational ecologists. Ecologists often want to characterize the distribution of vegetation concisely and quantitatively, as well as to assess the statistical significance of observed relationships [4]. In this example we will use a data set from the Bryce Canyon National Park in Utah. We will model the distribution of *Berberis (Mahonia) repens*, a common shrub of Bryce Canyon. Graphical analysis suggests that the distribution of the Berberis is related to elevation and possibly aspect value. Graphical data is shown in figure 1 below with the Berberis repens displayed in red. Appendix A serves the commands in R whereby we developed our results of the models in this chapter.

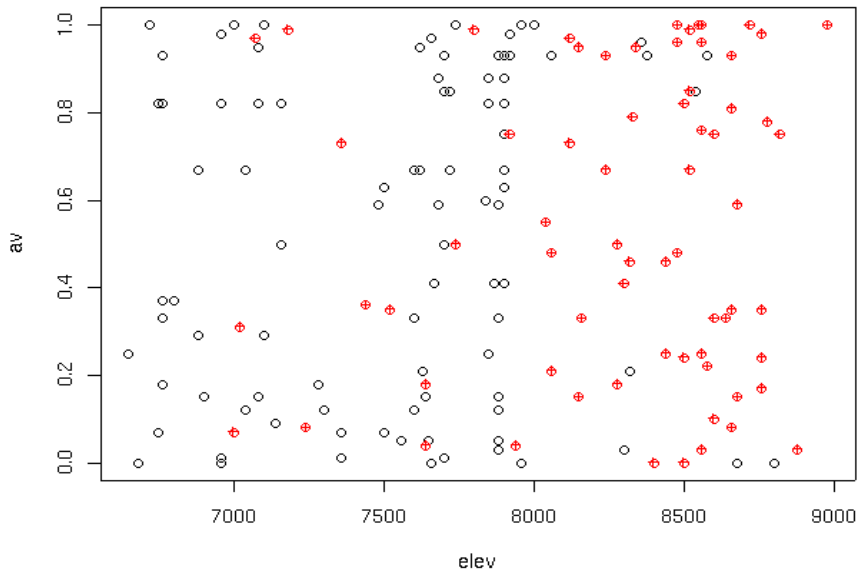

Figure 1: Aspect value relative to elevation of the Bryce Canyon data set.

3.1 Modeling Environment Relations with GLMs

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation [3]. In general the generalized linear model has a structure like

$$g(\mu_i) = \mathbf{X}_i\theta$$

where

$$\mu_i \equiv \mathbb{E}(Y_i) \quad \text{and} \quad Y_i \sim \text{some exponential family distribution}$$

Our g is a smooth monotonic function, $\mathbf{X}_i$ is a row of the model matrix for any strictly parametric model components, θ is the unknown parameter vector [8]. We can see from (1) that a GAM looks like a extensive version of a GLM.

This brief theoretical introduction about GLMs is an underlying support in understanding our computational results from R. The first graphical representation of the first order logistic fitted generalized linear model of the Berberis species is shown in figure 2.

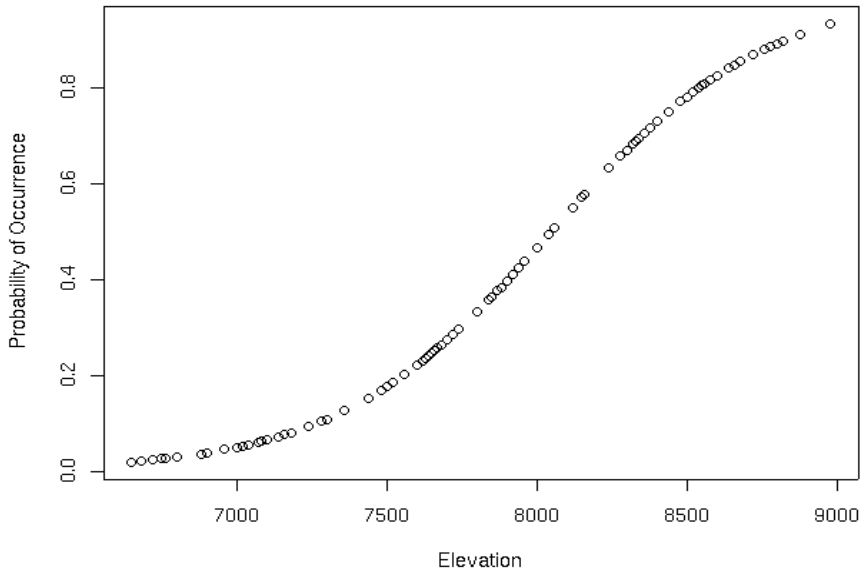

Figure 2: Generalized Linear Model of the Probability of Occurence relative to the Elevation.

Notice that the fitted curve is a smooth sigmoidal curve. The logistic regression is linear in the logit, but when back transformed from the logit to probabilities, it's sigmoidal. In addition, following conventional ecological theory we might assume that the probability exhibits a unimodal response to environment. To get a smooth unimodal response is simple. Just as a linear logistic regression is sigmoidal when back transformed to probability, a quadratic logistic regression is unimodal symmetric when back transformed to probabilities [4]. In figure 3 a GLM is shown that is quadratic for elevation. GLMs finesse the problems of bounded dependent variables and heterogeneous variances by transforming the dependent variable and employing a specific variance function for that transformation. While GLMs were shown to work reasonably well, and are in fact the method of choice for many ecologists, they are limited to linear functions [4].

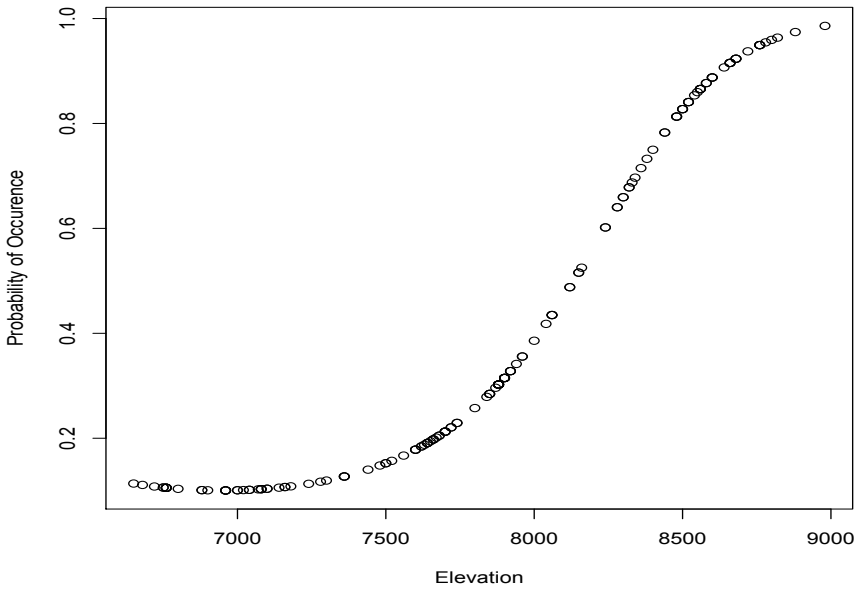

Figure 3: Generalized Linear Model of the Probability of Occurence relative to Quadratic Elevation.

When you examine the predicted values from GLMs, they are sigmoidal or modal curves, leading to the impression that they are not really linear.

3.2 Modeling Environment Relations with GAMs

Generalized Additive Models are designed to capitalize on the strengths of GLMs. In this case, the ability to fit logistic regressions without requiring the problematic steps of estimation of response curve shape or a specific parametric response function. As a practical matter, we can view GAMs as non-parametric curve fitters that attempt to achieve an optimal compromise between goodness-of-fit and parsimony of the final curve. Similar to GLMs, on species data they operate on deviance, rather than variance, and attempt to achieve the minimal residual deviance on the fewest degrees of freedom. The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called number of degrees of freedom. Also explained as; the number of observations minus the number of necessary relations among these observations. [7] One of the interesting aspects of GAMs is that they can only approximate the appropriate number of degrees of freedom, and that the number of degrees of freedom is often not an integer, but rather a real number with some fractional component.

We develop a GAM-model of the presence/absence of Berberis as a smooth function of elevation. The important elements from the computational analysis are the reduction in deviance for the degrees of freedom used. To get a visualisation

of what the result looks like, we made a plot seen in figure 4. In this GAM-model we have a reduction of the null deviance from 218.1935 to 133.4338 for 4 degrees of freedom. Given by R, the probability of achieving such a reduction is about 0.0004. Which proves a highly signifact lack of evidence to support the null hypothesis that logistic regression model provides an adequate fit for the data.

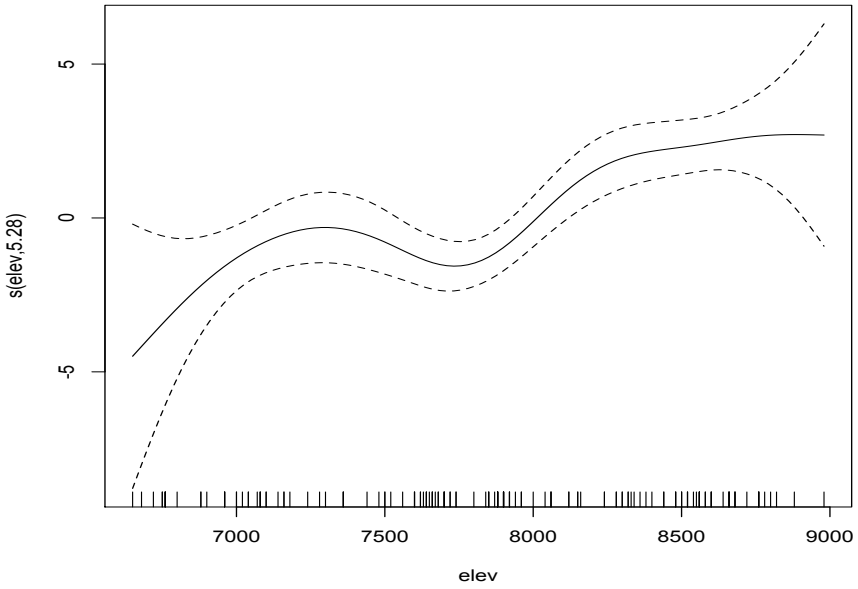

Figure 4: Model of the presence/absence of Berberis as a smooth function of elevation.

The default plot shows several things. The solid line is the predicted value of the dependent variable as a function of the elevation. The small lines along the x axis are the "rug", showing the location of the sample plots. The dashed lines are two times the standard errors of the estimates. The y axis is in the linear units, which in this case is logits, so that the values are centered on 0, and extend to both positive and negative values. To see the predicted values on the probability scale we need to use the back transformed values, which are shown in figure 5. Notice how the curve has multiple inflection points and modes, even though we did not specify a high order function. This is the beautiful property of GAMs. In order to fit the data, the function fit a bimodal curve. It seems unlikely ecologically that a species would have a multi-modal response to a single variable. Rather, it would appear to suggest competitive displacement by some other species from 7300 to 8000 feet, or the effects of another variable that interacts strongly over that elevation range [4].

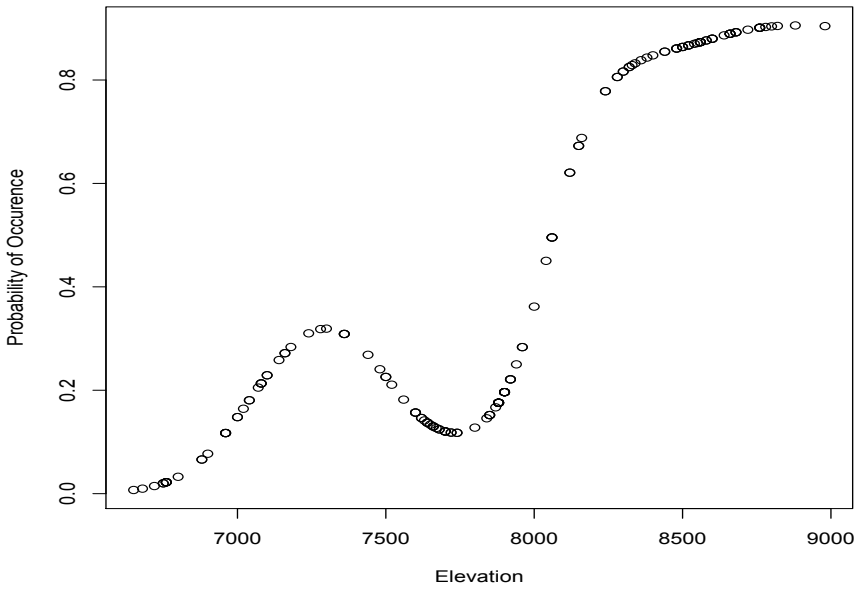

Figure 5: Model of the presence/absence of Berberis as a smooth function of elevation.

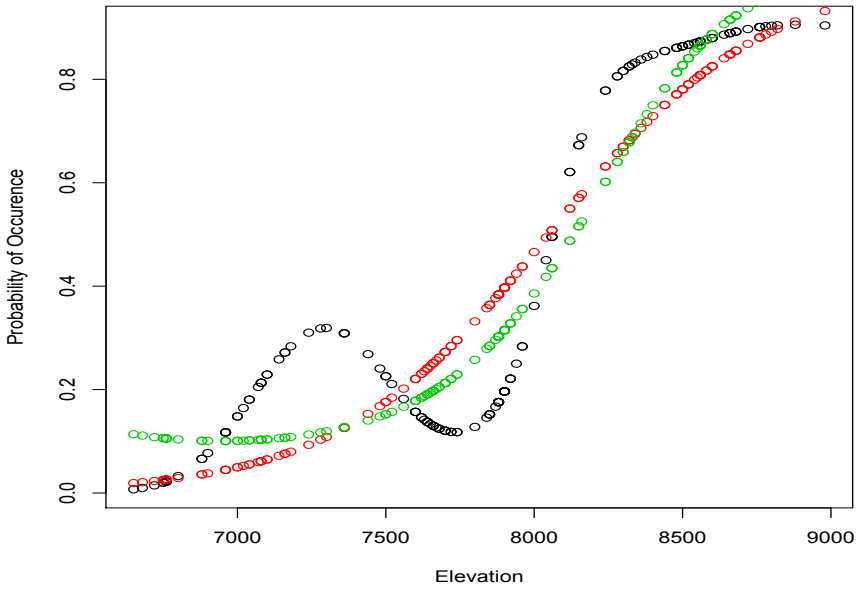

Figure 6: GLMs versus GAM. Predicted values from the first order logistic GLM in red, and the quadratic GLM in green

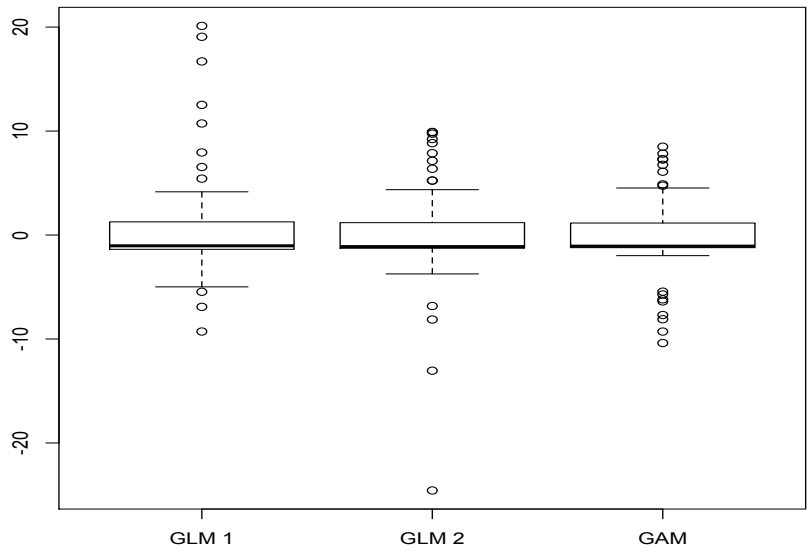

Figure 7: Boxplot of GLM1 (first order logistic GLM), GLM2 (quadratic logistic GLM) and the GAM.

3.3 GLMs versus GAMs

Figure 6 is the comparison of our last figure with the two GLM-models. It is the predicted values from the first order logistic GLM in red, and the quadratic GLM in green. Notice how even though the GAM curve is quite flexible, it avoids the problematic upturn at low values shown by the quadratic GLM. The GAM function results in a residual deviance of 129.80 on 153.72 degrees of freedom. Table 1 shows the difference of the GAM function to the GLMs.

model	**residual deviance**	**residual degrees of freedom**
GAM	129.80	153.72
linear GLM	151.95	158
quadratic GLM	146.63	157

Table 1: Residual deviance and residual degrees of freedom of GLM's and GAM.

For 3.28 degrees of freedom the GAM function earns a lower residual deviance of 16.83 compared to the quadratic GLM. The probability of achieving such a reduction in deviance in a nested model is 0.001. This is a highly significant result and while this is not strictly correct for a non-nested test, we conclude that our GAM-function is a better fit to the data. The AIC-values which is a value for the correctness of fit, also agrees with this statement. One approach

to a simple evaluation is to compare the residuals in a boxplot. This boxplot is given in figure 7. The three datasets are approximately balanced around zero, evidently the mean in both cases is near zero. However there is slightly more variation in GLM1 relative to GLM2, and there is also slightly more variation in GLM2 relative to GAM. The plot shows that the primary difference among the models is in the number of extreme residuals.

4 GAM Application on a Medical Article

In this chapter we will test the statistical analyses of a medical article with GAMs. The title of the article reads 'Are Cardiac Autonomic Nervous System Activity and Perceived Stress Related to Functional Somatic Symptoms in Adolescents? The TRAILS Study'. Stressors have been medically insufficiently explained related to functional somatic symptoms (FSS). However, the underlying mechanism of this association is largely unclear [2]. In the medical publication they examined whether FSS are associated with different perceived stress and cardiac autonomic nervous system (ANS) levels during a standardized stressful situation, and whether these associations are symptom-specific. Data from 715 adolescents from the Dutch cohort study Tracking Adolescents' Individual Lives Sample are collected during the Groningen Social Stress Test(GSST). FSS were clustered into a cluster of overtiredness, dizziness and musculoskeletal pain and a cluster of headache and gastrointestinal symp- toms. Perceived stress levels were splitted in unpleasantness and arousal stress, and cardiac ANS activity by assessing heart rate variability (HRV-HF) and pre-ejection period (PEP). Perceived stress and cardiac ANS levels before, during, and after the GSST were studied as well as cardiac ANS reactivity [2]. In the medical article the authors solely used linear regression analyses to examine the associations. The relation of stressors relative to FSS remain unclear by these linear analysis. Our main goal is to clarify the relations by using GAMs and demonstrate that medical data should be analysed more often with GAMs. For a more comprehensive explanation of the medical definitions and relationships in this chapter, please read the article found in the references. All analysis in this chapter were performed in R. All commands in R can be found in Appendix B.

Tables 2, 4, 6 and 8 in this chapter are partially designed in R to be a copy of table 2 and 3 in the article. The results from R have a small differance relative to the results in the article. The cause could be the intermediate rounding of numbers by the authors. Fortunately, the differance is minimal and insignifact and we take them therefore as negligible. Interpret the numbers given in table 2, 4, 6, and 8 as; Beta(p-value), likewise as the tables in the article. The figures in blue are the attached adjusted R^2-values.

Tables 3, 5, 7 and 9 are the results from the GAM in R. Interpret the results given in the odd-numbered tables as; degree-of-freedom(p-value) and the figure in blue as the adjusted R^2. Significance in all tables is designated by; $^{*}p < 0.05, ^{**}p < 0.01$.

4.1 Functional Somatic Symptoms Related to Perceived Stress

Table 2 and table 4 display the first linear regressions analysis in the article. These linear regression analyses were to examine whether functional somatic symptoms (FSS) were related to perceived stress before, during or after the GSST. Both table results are adjusted for sex and medical use.

Compare table 2 with table 3. Notice that whenever the degree of freedom of the GAM is 1 we get the same adjusted R^2 and p-value as our linear model.

This is as expected since a GAM is the same as a linear model for this degree of freedom. This can be seen in figures 8 and 9.

We found three results by the GAM with a lower p-value compared with the linear model. Two of these remain insignificant while one of the linear models was significant and is highly signficant by the GAM. Notable is that the result of a GAM is not nessecarry better compared with a linear model(e.g. where 'perceived arousal after' is significant to 'FSS' in the linear model, it is not by the GAM). The adjusted R^2 values of the GAMs are always higher or equal to those of the linear models. Whenever the value is higher, it means that the GAM gives a better fitted regression line to the data. By the results we notice that the GAM adjusts its model by the adjusted R^2 value. Negative adjusted R^2 values could occur whenever the model is adjusted for variables that do not help to predict the response. Figures 8, 9 and 10 show a graphical representation of the GAMs. Another three lower p-values are found by the GAM in table 5. Unfortunately, none of these show a signifant results.

	Perceived arousal before	**Perceived arousal during**	**Perceived arousal after**
Functional somatic symptoms	0.06(0.11) −0.000357	0.10(0.011)* 0.00785	0.08(0.046)* 0.00392
Cluster of headache and gastrointestinal symptoms	−0.01(0.75) −0.00384	0.05(0.22) 0.00652	−0.06(0.18) −0.0020
Cluster of over-tiredness, dizziness and musculoskeletal pain	0.09(0.042)* 0.001495	0.10(0.023)* 0.00998	0.16(0.003)** 0.0156

Table 2: Perceived arousal stress associations by the linear model.

	Perceived arousal before	**Perceived arousal during**	**Perceived arousal after**
Functional somatic symptoms	1(0.11) −0.000361	3.664(0.001)** 0.0279	1.663(0.07) 0.00631
Cluster of headache and gastrointestinal symptoms	1(0.75) −0.00384	4.335(0.17) 0.0153	1.94(0.27) −0.0020
Cluster of over-tiredness, dizziness and musculoskeletal pain	3.535(0.07) 0.00615	1(0.027)* 0.00996	4.45(0.002)** 0.0256

Table 3: Perceived arousal stress associations by the generalized additive model.

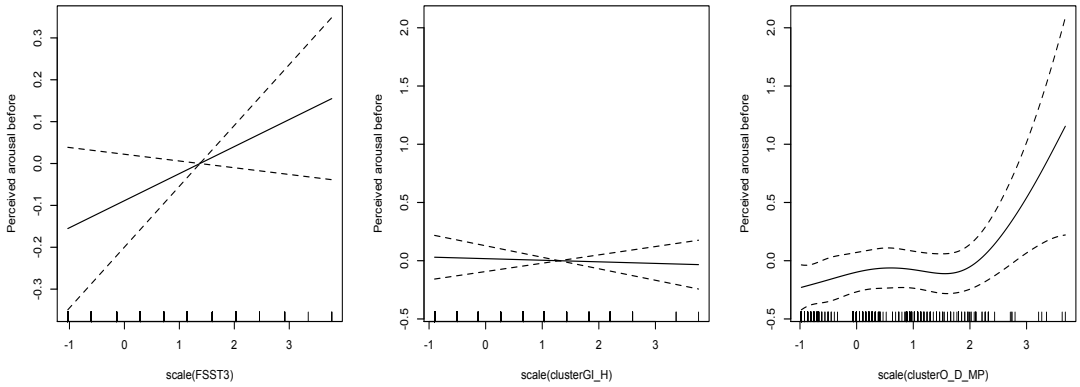

Figure 8: Associations of Perceived Arousal Before the GSST.

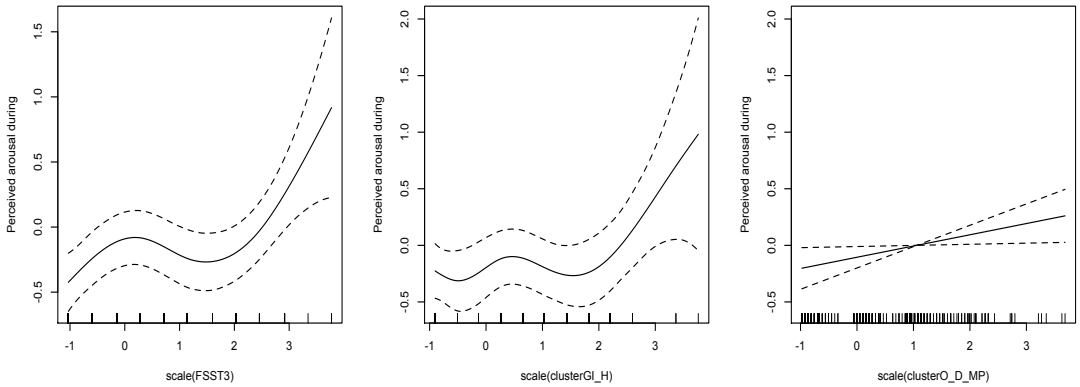

Figure 9: Associations of Perceived Arousal During the GSST.

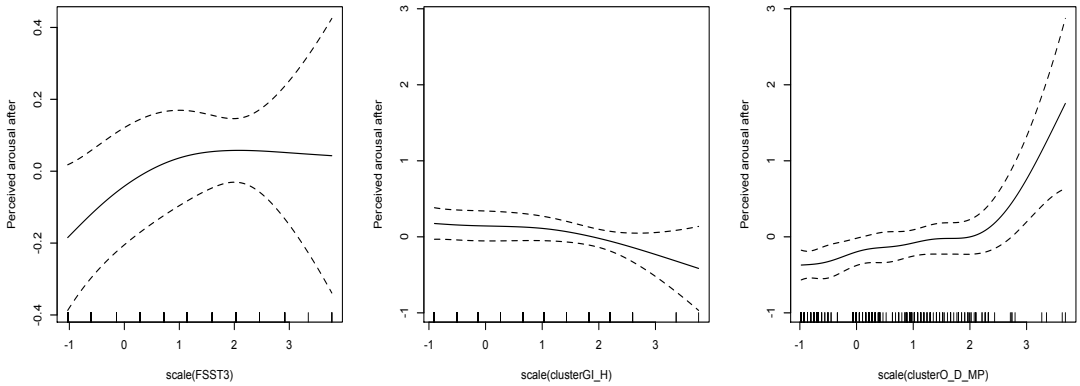

Figure 10: Associations of Perceived Arousal After the GSST.

	Perceived unpleasantness before	**Perceived unpleasantness during**	**Perceived unpleasantness after**
Functional somatic symptoms	$0.08(0.047)^{*}$ 0.002596	$0.14(0.0007)^{**}$ 0.01339	$0.07(0.11)$ 0.005586
Cluster of headache and gastrointestinal symptoms	$-0.02(0.60)$ -0.00239	$0.09(0.03)^{*}$ 0.00984	$0.03(0.44)$ 0.00418
Cluster of overtiredness, dizziness and musculoskeletal pain	$0.13(0.005)^{**}$ 0.0050	$0.07(0.10)$ 0.00612	$0.04(0.40)$ 0.00522

Table 4: Perceived unpleasantness stress associations by the linear model.

	Perceived unpleasantness before	**Perceived unpleasantness during**	**Perceived unpleasantness after**
Functional somatic symptoms	$2.241(0.06)$ 0.00813	$3.152(0.002)^{**}$ 0.0205	$2.095(0.057)$ 0.013
Cluster of headache and gastrointestinal symptoms	$1(0.61)$ -0.00239	$1(0.03)^{*}$ 0.00984	$1.74(0.34)$ 0.00733
Cluster of overtiredness, dizziness and musculoskeletal pain	$1(0.005)^{**}$ 0.0052	$4.077(0.27)$ 0.0124	$1(0.33)$ 0.00522

Table 5: Perceived unpleasantness stress associations by the generalized additive model.

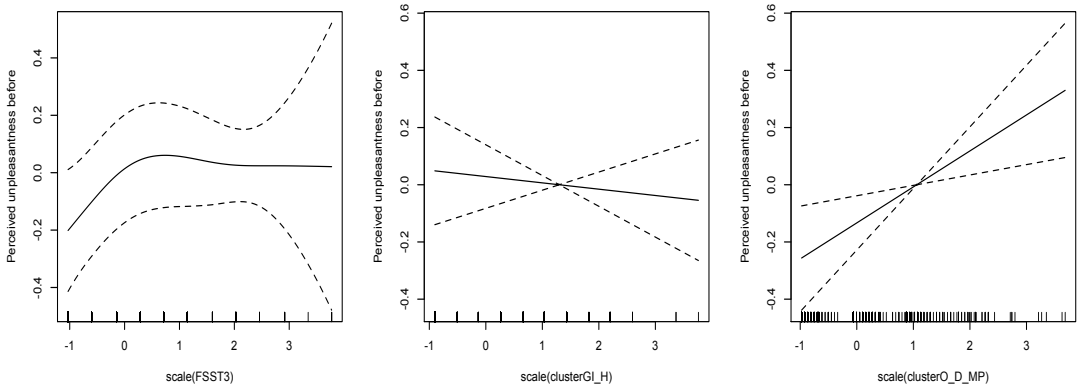

Figure 11: Associations of Perceived Unpleasantness Before the GSST.

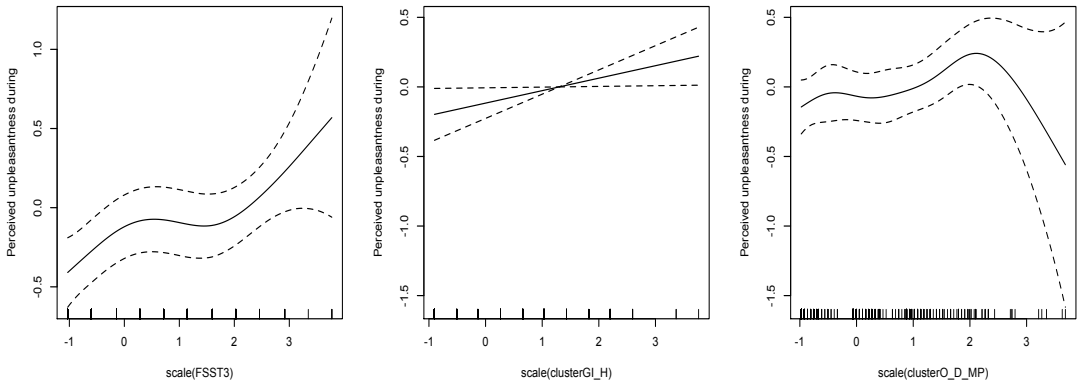

Figure 12: Associations of Perceived Unpleasantness During the GSST.

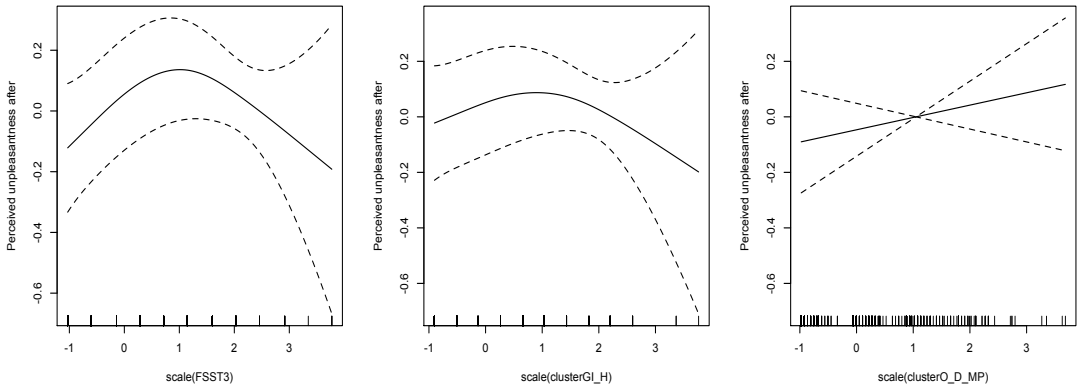

Figure 13: Associations of Perceived Unpleasantness After the GSST.

4.2 Functional Somatic Symptoms Related to Heart Rate Variability and Pre-Ejection Period

Another 18 linear regression analyses were performed to examine whether FSS were related to heart rate variability (HRV-HF) or pre-ejection period (PEP) before, during or after the GSST. As shown in table 6 and 8 no significant associations were found by the linear regressions analysis. The GAMs also had an insignificant result unfortunately. The GAMs showed many similarities to the linear models and did not result in new significant associations. Therefore we can conclude that the analysis performed on this article are not significantly improved by using the GAM.

	HRV-HF before	**HRV-HF during**	**HRV-HF after**
Functional somatic symptoms	0.02(0.64) 0.0177	−0.004(0.94) 0.0404	0.02(0.65) 0.0186
Cluster of headache and gastrointestinal symptoms	0.001(0.99) 0.0158	−0.06(0.17) 0.0424	0.03(0.48) 0.0184
Cluster of overtiredness, dizziness and musculoskeletal pain	0.01(0.81) 0.0133	0.06(0.25) 0.0486	−0.02(0.63) 0.018

Table 6: HRV-HF associations by the linear model.

	HRV-HF before	**HRV-HF during**	**HRV-HF after**
Functional somatic symptoms	1.65(0.57) 0.0197	1(0.94) 0.0404	1(0.66) 0.0186
Cluster of headache and gastrointestinal symptoms	1(0.98) 0.0160	1(0.17) 0.0424	1(0.48) 0.0184
Cluster of overtiredness, dizziness and musculoskeletal pain	1.22(0.84) 0.0133	1(0.25) 0.0486	1(0.63) 0.018

Table 7: HRV-HF associations by the generalized additive model.

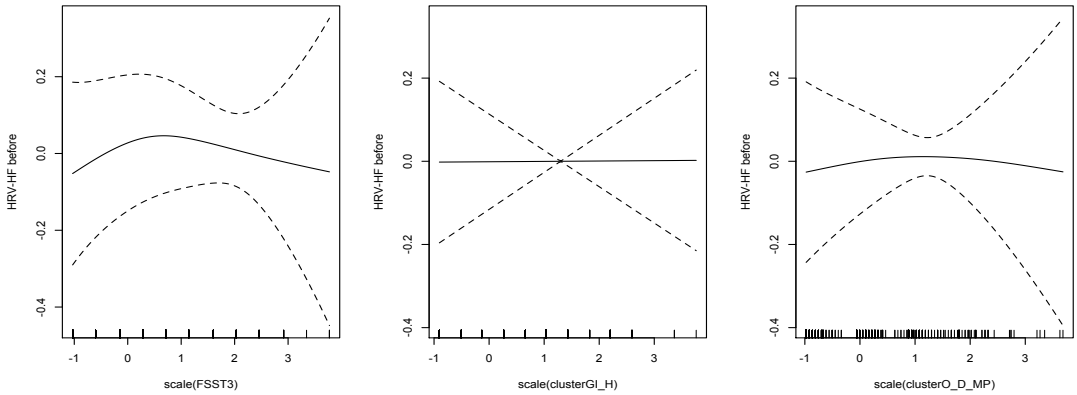

Figure 14: Associations of HRV-HF Before the GSST.

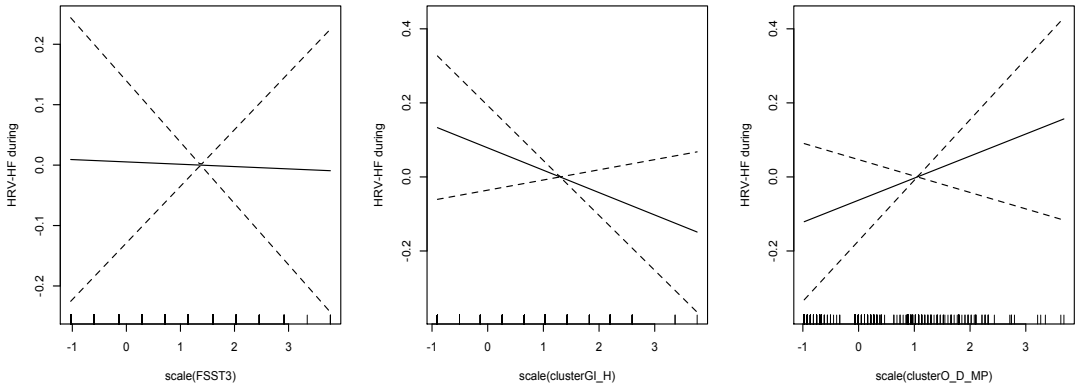

Figure 15: Associations of HRV-HF During the GSST.

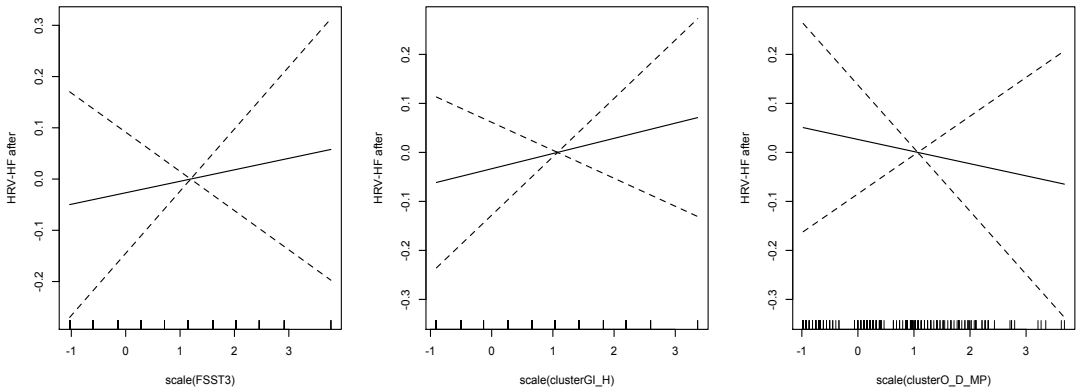

Figure 16: Associations of HRV-HF After the GSST.

	PEP before	**PEP during**	**PEP after**
Functional somatic symptoms	0.05(0.32) 0.0397	0.07(0.19) 0.0428	0.04(0.48) 0.0348
Cluster of headache and gastrointestinal symptoms	−0.04(0.46) 0.0375	0.000(1.00) 0.0387	−0.04(0.47) 0.0322
Cluster of over-tiredness, dizziness and musculoskeletal pain	0.11(0.05) 0.0462	0.07(0.26) 0.0316	0.08(0.22) 0.0326

Table 8: PEP associations by the linear model.

	PEP before	**PEP during**	**PEP after**
Functional somatic symptoms	1(0.32) 0.0397	3.37(0.31) 0.0478	1(0.48) 0.0348
Cluster of headache and gastrointestinal symptoms	2.62(0.47) 0.0426	3.10(0.56) 0.0434	2.90(0.45) 0.0377
Cluster of over-tiredness, dizziness and musculoskeletal pain	1(0.07) 0.0462	1(0.28) 0.0316	1(0.22) 0.0326

Table 9: PEP associations by the generalized additive model.

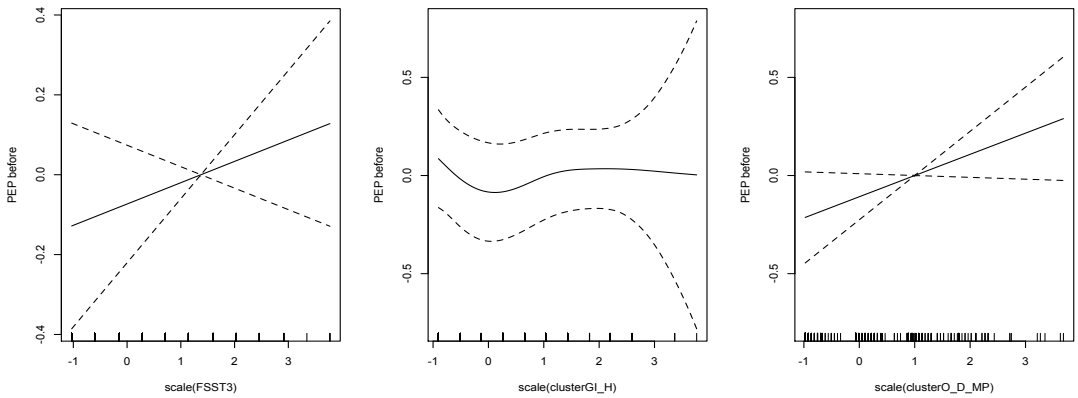

Figure 17: Associations of PEP Before the GSST.

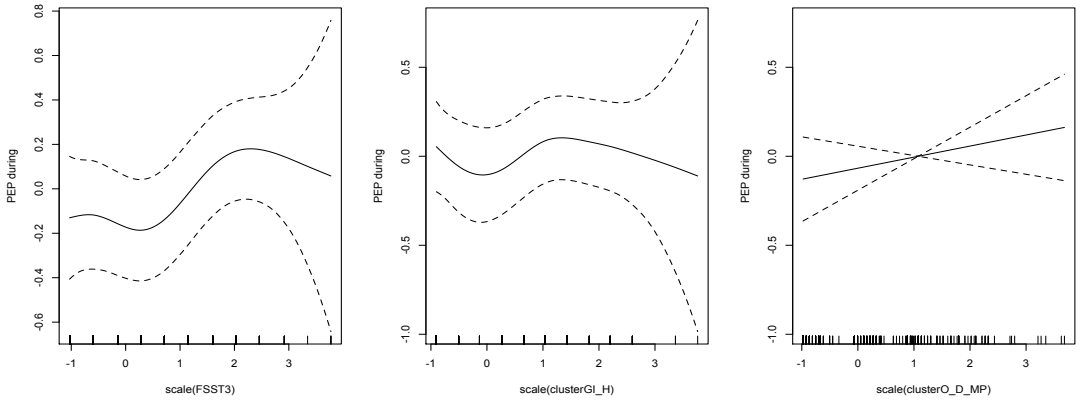

Figure 18: Associations of PEP During the GSST.

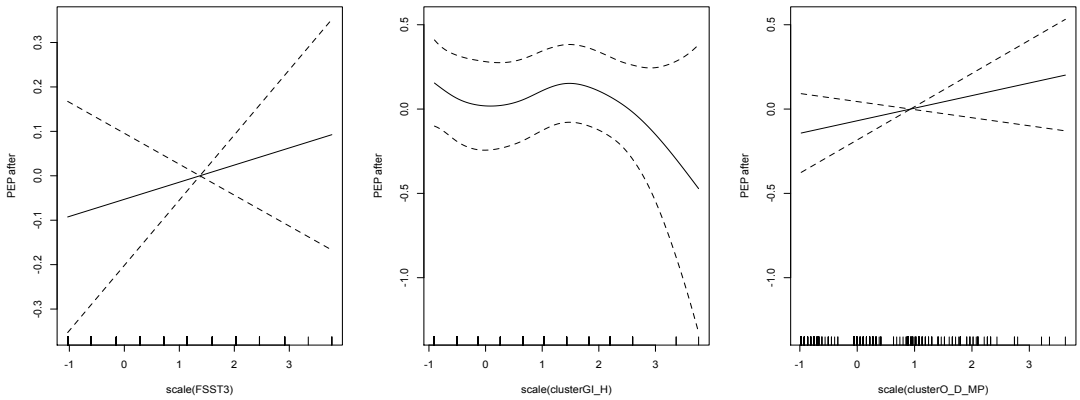

Figure 19: Associations of PEP After the GSST.

5 Epilogue

In this thesis we started with the theory and properties of generalized additive models. Succeeded in chapter 3 by an example to get a better visualization of our subject. This was preparatory work for our main goal in chapter 4; application of the theory on a medical article. As a result of the outcome we suppose that the GAM shapes its model by the adjusted R^2-value. These values always showed an improvement whenever the GAM did not result in a linear regression spline. Despite the adjusted R^2-values, we have seen that the GAMs did not result in a new significant association relative to the linear models. On the contrary, some figures do show non-linearity by the GAM. This non-linearity could even increase whenever we would increase the sample size and gain more extreme values. Therefore, we still hypothesize that GAMs could improve medical analysis.

Much of the inserted time in this thesis was necessary to find an medical article with its dataset. The transparancy of medical data is very limited with the cause of privacy surveillance. This is an obstruction for science while most datasets only consists of numbers. In this matter, we have examined three medical publications and two of which were inadequate interesting to test. The article we have examined in this thesis has been worthwhile but unfortunately the results were insignificant. No time was left to find new medical articles which could have resulted in a more interesting outcome.

References

[1] C. Gu. Smoothing spline anova models. 2002.

[2] A.M. Van Roon J.A.M. Hunfeld P.F.C. Groot A.J. Oldehinkel J.G.M. Rosmalen K.A.M. Janssens, H. Riese1. Are cardiac autonomic nervous system activity and perceived stress related to functional somatic symptoms in adolescents? the trails study. *PLOS one*, 2016.

[3] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):384, 1972.

[4] D.W. Roberts. R labs for vegetation ecologists, 2016.

[5] R.J. Tibshirani T.J. Hastie. Generalized additive models. 1990.

[6] G. Wahba. Spline models for observational data. 1990.

[7] H.M. Walker. Degrees of freedom. *Journal of Educational Psychology*, 31(4):253–269, 1940.

[8] Simon N. Wood. Generalized additive models: An introduction with r. *Statistical Science*, 2006.

Appendix A

```
# EXAMPLE Bryce Canyon data set
library(mgcv)
#Datasets
veg = read.table("/Users/Teyler/Desktop/Scriptie/bryceveg.R",
header = TRUE, row.names=1)
site = read.table("/Users/Teyler/Desktop/Scriptie/brycesite.R",
header = TRUE, row.names=1)
attach(site)

#Probability of occurence versus 1st order elevation
bere1_glm <- glm (veg$berrep>0 ~ elev, family = binomial)
plot(elev,fitted(bere1_glm),xlab='Elevation', ylab=
'Probability of Occurrence')

#Probability of occurence versus quadratic elevation
bere3_glm<-glm(veg$berrep>0~elev+I(elev^2),family=binomial)
plot(elev,fitted(bere3_glm), xlab='Elevation', ylab=
'Probability of Occurence')

#GAMs
bere1_gam <- gam(veg$berrep>0~s(elev),family=binomial)
plot(bere1_gam)
plot(elev,fitted(bere1_gam), xlab='Elevation', ylab=
'Probability of Occurence')

#GAMs vs. GLMs
plot(elev,fitted(bere1_gam), xlab='Elevation', ylab=
'Probability of Occurence')
points(elev,fitted(bere1_glm),col=2)
points(elev,fitted(bere3_glm),col=3)

anova(bere3_glm,bere1_gam,test='Chi')
boxplot(bere1_glm$residuals,bere3_glm$residuals,
bere1_gam$residuals,names=c("GLM 1","GLM 2","GAM"))
```

Appendix B

```
#clear previous objects
rm(list = ls())
#Loading packages for function commands
library(mgcv)
#All empty spaces in excel are replaced by the value 999
mydata = read.csv2("/Users/Teyler/Desktop/Scriptie/Publications/
2/S1Dataset999.csv", header = TRUE)
#Replace our outlying values with 'Not␣Available␣(NA)'
mydata[mydata==999] <- NA
#Check if R is importing our excel file properly
str(mydata)

#Checking table 1. Sample characteristics with 'x' the column we
want to check
length(mydata[,x]) - sum(is.na(mydata[,x]))
mean(mydata[,x], na.rm = TRUE)
sd(mydata[,x], na.rm = TRUE)

#LM analysis on first row(FSS) Table 2. The associations between
perceived stress before, during, and after the GSST task and FSS
lm1ab<- lm(scale(Arousal_before) ~ scale(FSST3) + sex + T3med,
data=mydata, weights=mydata[,9])
lm1ad<- lm(scale(Arousal_stress) ~ scale(FSST3) + sex + T3med,
data=mydata, weights=mydata[,9])
lm1aa<- lm(scale(Arousal_after) ~ scale(FSST3) + sex + T3med,
data=mydata, weights=mydata[,9])
lm1ub<- lm(scale(Unpleasantness_before) ~ scale(FSST3) + sex +
T3med, data=mydata, weights=mydata[,9])
lm1ud<- lm(scale(Unpleasantness_stress) ~ scale(FSST3) + sex +
T3med, data=mydata, weights=mydata[,9])
lm1ua<- lm(scale(Unpleasantness_after) ~ scale(FSST3) + sex +
T3med, data=mydata, weights=mydata[,9])
summary(lm1ab)
summary(lm1ad)
summary(lm1aa)
summary(lm1ub)
summary(lm1ud)
summary(lm1ua)

#LM analysis on second and third row Table 2. The associations
between perceived stress before, during, and after the GSST task
and FSS
lm2ab<- lm(scale(Arousal_before) ~ scale(clusterGI_H) +
scale(clusterO_D_MP) +sex + T3med, data=mydata, weights=mydata[,9])
lm2ad<- lm(scale(Arousal_stress) ~ scale(clusterGI_H) +
scale(clusterO_D_MP) +sex + T3med, data=mydata, weights=mydata[,9])
lm2aa<- lm(scale(Arousal_after) ~ scale(clusterGI_H) +
```

```
scale(clusterO_D_MP) +sex + T3med, data=mydata, weights=mydata[,9])
lm2ub<- lm(scale(Unpleasantness_before) ~ scale(clusterGI_H) +
scale(clusterO_D_MP) +sex + T3med, data=mydata, weights=mydata[,9])
lm2ud<- lm(scale(Unpleasantness_stress) ~ scale(clusterGI_H) +
scale(clusterO_D_MP) +sex + T3med, data=mydata, weights=mydata[,9])
lm2ua<- lm(scale(Unpleasantness_after) ~ scale(clusterGI_H) +
scale(clusterO_D_MP) +sex + T3med, data=mydata, weights=mydata[,9])
summary(lm2ab)
summary(lm2ad)
summary(lm2aa)
summary(lm2ub)
summary(lm2ud)
summary(lm2ua)

#GAM analysis on first row(FSS) Table 2. The associations
between perceived stress before, during, and after the GSST
task and FSS
gam1ab <- gam(scale(Arousal_before) ~ s(scale(FSST3)) + sex
+ T3med, data=mydata, weights = mydata[,9])
gam1ad <- gam(scale(Arousal_stress) ~ s(scale(FSST3)) + sex
+ T3med, data=mydata, weights = mydata[,9])
gam1aa <- gam(scale(Arousal_after) ~ s(scale(FSST3)) + sex
+ T3med, data=mydata, weights = mydata[,9])
gam1ub <- gam(scale(Unpleasantness_before) ~ s(scale(FSST3))
+ sex + T3med, data=mydata, weights = mydata[,9])
gam1ud <- gam(scale(Unpleasantness_stress) ~ s(scale(FSST3))
+ sex + T3med, data=mydata, weights = mydata[,9])
gam1ua <- gam(scale(Unpleasantness_after) ~ s(scale(FSST3))
+ sex + T3med, data=mydata, weights = mydata[,9])
summary(gam1ab)
summary(gam1ad)
summary(gam1aa)
summary(gam1ub)
summary(gam1ud)
summary(gam1ua)

#GAM Analysis on second and third row Table 2. The associations
between perceived stress before, during, and after the GSST task
and FSS
gam2ab<- gam(scale(Arousal_before) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) +sex + T3med, data=mydata, weights=mydata[,9])
gam2ad<- gam(scale(Arousal_stress) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) +sex + T3med, data=mydata, weights=mydata[,9])
gam2aa<- gam(scale(Arousal_after) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) +sex + T3med, data=mydata, weights=mydata[,9])
gam2ub<- gam(scale(Unpleasantness_before) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) +sex + T3med, data=mydata, weights=mydata[,9])
gam2ud<- gam(scale(Unpleasantness_stress) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) +sex + T3med, data=mydata, weights=mydata[,9])
gam2ua<- gam(scale(Unpleasantness_after) ~ s(scale(clusterGI_H)) +
```

```
s(scale(cluster0_D_MP)) +sex + T3med, data=mydata, weights=mydata[,9])
summary(gam2ab)
summary(gam2ad)
summary(gam2aa)
summary(gam2ub)
summary(gam2ud)
summary(gam2ua)

#Combiplots gam1/gam2
par(mfrow=c(1,3))
plot(gam1ab, ylab = 'Perceived␣arousal␣before')
plot(gam2ab, ylab = 'Perceived␣arousal␣before')
par(mfrow=c(1,3))
plot(gam1ad, ylab = 'Perceived␣arousal␣during')
plot(gam2ad, ylab = 'Perceived␣arousal␣during')
par(mfrow=c(1,3))
plot(gam1aa, ylab = 'Perceived␣arousal␣after')
plot(gam2aa, ylab = 'Perceived␣arousal␣after')

par(mfrow=c(1,3))
plot(gam1ub, ylab = 'Perceived␣unpleasantness␣before')
plot(gam2ub, ylab = 'Perceived␣unpleasantness␣before')
par(mfrow=c(1,3))
plot(gam1ud, ylab = 'Perceived␣unpleasantness␣during')
plot(gam2ud, ylab = 'Perceived␣unpleasantness␣during')
par(mfrow=c(1,3))
plot(gam1ua, ylab = 'Perceived␣unpleasantness␣after')
plot(gam2ua, ylab = 'Perceived␣unpleasantness␣after')

#Analysis on variance testing (ANOVA)
anova(gam1ab, lm1ab)
with(mydata, plot(Arousal_before, FSST3))


#LM analysis on first row (FSS) Table 3. The associations
between ANS activity before, during, and after the GSST and
FSS assesed in sitting position
lm1hfb<- lm(scale(HF_HRV_before) ~ scale(FSST3) + sex + T3med +
Smoking + Exercise + bmi + dept3, data=mydata, weights=mydata[,9])
lm1hfd<- lm(scale(HF_HRV_stress) ~ scale(FSST3) + sex + T3med +
Smoking + Exercise + bmi + dept3, data=mydata, weights=mydata[,9])
lm1hfa<- lm(scale(HF_HRV_after) ~ scale(FSST3) + sex + T3med +
Smoking + Exercise + bmi + dept3, data=mydata, weights=mydata[,9])
lm1pepb<- lm(scale(PEP_before) ~ scale(FSST3) + sex + T3med +
Smoking + Exercise + bmi + dept3, data=mydata, weights=mydata[,9])
lm1pepd<- lm(scale(PEP_stress) ~ scale(FSST3) + sex + T3med +
Smoking + Exercise + bmi + dept3, data=mydata, weights=mydata[,9])
lm1pepa<- lm(scale(PEP_after) ~ scale(FSST3) + sex + T3med +
Smoking + Exercise + bmi + dept3, data=mydata, weights=mydata[,9])
summary(lm1hfb)
```

```
summary(lm1hfd)
summary(lm1hfa)
summary(lm1pepb)
summary(lm1pepd)
summary(lm1pepa)

#LM analysis on second and third row Table 3. The associations
between ANS activity before, during, and after the GSST and
FSS assesed in sitting position
lm2hfb<- lm(scale(HF_HRV_before) ~ scale(clusterGI_H) +
scale(cluster0_D_MP) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
lm2hfd<- lm(scale(HF_HRV_stress) ~ scale(clusterGI_H) +
scale(cluster0_D_MP) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
lm2hfa<- lm(scale(HF_HRV_after) ~ scale(clusterGI_H) +
scale(cluster0_D_MP) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
lm2pepb<- lm(scale(PEP_before) ~ scale(clusterGI_H) +
scale(cluster0_D_MP) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
lm2pepd<- lm(scale(PEP_stress) ~ scale(clusterGI_H) +
scale(cluster0_D_MP) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
lm2pepa<- lm(scale(PEP_after) ~ scale(clusterGI_H) +
scale(cluster0_D_MP) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
summary(lm2hfb)
summary(lm2hfd)
summary(lm2hfa)
summary(lm2pepb)
summary(lm2pepd)
summary(lm2pepa)

#GAM analysis on first row (FSS) Table 3. The associations
between ANS activity before, during, and after the GSST and
FSS assesed in sitting position
gam1hfb<- gam(scale(HF_HRV_before) ~ s(scale(FSST3)) + sex
+ T3med + Smoking + Exercise + bmi + dept3, data=mydata,
weights=mydata[,9])
gam1hfd<- gam(scale(HF_HRV_stress) ~ s(scale(FSST3)) + sex
+ T3med + Smoking + Exercise + bmi + dept3, data=mydata,
weights=mydata[,9])
gam1hfa<- gam(scale(HF_HRV_after) ~ s(scale(FSST3)) + sex
+ T3med + Smoking + Exercise + bmi + dept3, data=mydata,
weights=mydata[,9])
gam1pepb<- gam(scale(PEP_before) ~ s(scale(FSST3)) + sex
+ T3med + Smoking + Exercise + bmi + dept3, data=mydata,
weights=mydata[,9])
gam1pepd<- gam(scale(PEP_stress) ~ s(scale(FSST3)) + sex
```

```
+ T3med + Smoking + Exercise + bmi + dept3, data=mydata,
weights=mydata[,9])
gam1pepa<- gam(scale(PEP_after) ~ s(scale(FSST3)) + sex
+ T3med + Smoking + Exercise + bmi + dept3, data=mydata,
weights=mydata[,9])
summary(gam1hfb)
summary(gam1hfd)
summary(gam1hfa)
summary(gam1pepb)
summary(gam1pepd)
summary(gam1pepa)

#GAM analysis on second and third row Table 3. The associations
between ANS activity before, during, and after the GSST and FSS
assesed in sitting position
gam2hfb<- gam(scale(HF_HRV_before) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
gam2hfd<- gam(scale(HF_HRV_stress) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
gam2hfa<- gam(scale(HF_HRV_after) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
gam2pepb<- gam(scale(PEP_before) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
gam2pepd<- gam(scale(PEP_stress) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
gam2pepa<- gam(scale(PEP_after) ~ s(scale(clusterGI_H)) +
s(scale(clusterO_D_MP)) + sex + T3med + Smoking + Exercise +
bmi + dept3, data=mydata, weights=mydata[,9])
summary(gam2hfb)
summary(gam2hfd)
summary(gam2hfa)
summary(gam2pepb)
summary(gam2pepd)
summary(gam2pepa)

#Combiplots table 3 gam1/gam2
par(mfrow=c(1,3))
plot(gam1hfb, ylab = 'HRV-HF␣before')
plot(gam2hfb, ylab = 'HRV-HF␣before')
par(mfrow=c(1,3))
plot(gam1hfd, ylab = 'HRV-HF␣during')
plot(gam2hfd, ylab = 'HRV-HF␣during')
par(mfrow=c(1,3))
plot(gam1hfa, ylab = 'HRV-HF␣after')
plot(gam2hfa, ylab = 'HRV-HF␣after')
```

```
par(mfrow=c(1,3))
plot(gam1pepb, ylab = 'PEP before')
plot(gam2pepb, ylab = 'PEP before')
par(mfrow=c(1,3))
plot(gam1pepd, ylab = 'PEP during')
plot(gam2pepd, ylab = 'PEP during')
par(mfrow=c(1,3))
plot(gam1pepa, ylab = 'PEP after')
plot(gam2pepa, ylab = 'PEP after')
```