HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

Literature Review of the Generalised Additive Model for location, scale and shape

Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Master in Science (M.Sc) in
Statistics and Operations Research

Author
Konstantinos Zacharias

Supervisor
Loukia Meligotsidou

Department of Mathematics
Athens, August 2021

This thesis includes information acquired from authentic and highly regarded resources. Reasonable efforts have been made to publish reliable material. However, the author and writer cannot assume responsibility for the validity of all resources and the repercussions of their use. There has been a major attempt in tracking the copyright holders of all materials reported in this dissertaition and sincerely apologize to any copyright holders if their permission has not been obtained.

First printing, 2021.

To my mother, Aristea.

Acknowledgements

The fulfillment of this dissertation comes as the last milestone to the completion of this postgraduate program. It has been a period of life acquiring priceless experiences and altering my train of thought. I feel the urging need to express my gratitude for all of my teachers who were devoted and passionate and motivated me until the end.

I owe a deep sense of gratitude to Prof. Fotios Sianis and Prof. Samis Trevezas who willingly pertained in this evaluation committee. This particular thesis would not have been written without their continuous supervision and guidance.

It is apparent that writing this scientific dissertation would have been impossible, if it was not the sincere contribution of my supervisor teacher Prof. Loukia Meligotsidou, who endorsed and guided me with her valuable knowledge and expertise constantly. I am excited that I had the chance to cooperate with such a bright statistician who motivated me from the very beginning and I was extremely impressed by her ingenious personality and the degree of freedom she allows to her students to find their own stepping and mature as new scientists.

My deepest appreciation and gratitude is attributed to all of my colleagues and friends who encouraged and supported me to go on through the toughest issues I faced.

It is essential to express my special gratitude towards my family for the ongoing support and encouragement which assisted me to overcome the hurdles and ultimately complete my scientific piece of paper.

Abstract

The main objective of this study is to present different statistical models and discuss their contribution to data fit. The first model that is analysed is the Generalised Linear Model(GLM) which is a generalisation of the linear model assuming for the distribution of the response variable to be a member of the exponential family of distributions. The nature of the data determines to a great extent the form of the generalised linear model that will be applied, through the choice of the link function of the model. The iterative methods which allow for the practical implementation of each particular model and the respective statistical inference procedures are discussed, as well.

The assumption of the exponential family distribution for the response holds in the Generalised Additive Model(GAM). A distribution that belongs in the exponential family of distributions is assumed for the dependent variable, with the introduction of smoothing functions that blend the inherent properties of the GLM with the additive models. The response variable depends linearly on unknown smooth functions of some predictor variables, and the inference is focused on these smoothers.

A general class of statistical models for a univariate response variable is presented, which is called the Generalized Additive Model for Location, Scale and Shape (GAMLSS). The choice of the distribution for the response variable in GAMLSS is made from a very general family of distributions including highly skewed or kurtotic continuous and discrete distributions. The GAMLSS systematic part is expanded to permit modelling of the mean (or location) and other distributional parameters of the response, as parametric and/or additive non-parametric (smooth) functions of explanatory variables and/or random-effects terms.

Contents

List of Figures

Chapter 1

The Generalised Linear Model

The generalised linear model (GLM) extends the linear regression model as to achieve a better fit of the data whenever it is not feasible due to certain circumstances under which linear model is no longer efficient. Major issues that generalised linear models deal are non normal response distributions and the heteroscedasticity problem. The main structure of the generalised linear models is summarised in three model components; a random component, a systematic component and a link function. The random component identifies the response variable Y and its respective probability distribution, the systematic component of the generalised linear models clarifies the explanatory variables that consist the linear predictor functions, as for the link function it elaborates the function of average Y that the model is equal to the linear combination of the explanatory variables.

1.1 Parts of the Generalised Linear Model

In this section, a thorough look into the model components is taken. The random component of a generalised linear model is related with the probability distribution of the response variable Y. Let Y be the response variable with independent observations $(y_1, \ldots, y_N)$ from a distribution in the natural exponential family of the form

$$f(y_i; \theta_i, \phi) = exp((y_i\theta_i - b(\theta_i))/\alpha(\phi) + c(y_i, \phi)). \tag{1.1}$$

This is called the exponential dispersion family (Agresti, 2015).

The parameter θ_i is called the natural parameter, and ϕ is the dispersion parameter. If $\alpha(\phi) = 1,\ c(y_i, \phi) = c(y_i)$, the natural exponential family is derived

$$f(y_i; \theta_i) = \alpha(\theta_i)b(y_i)exp(y_iQ(\theta_i)). \tag{1.2}$$

Otherwise, usually $\alpha(\phi)$ is of the form $\alpha(\phi) = \phi$, or $\alpha(\phi) = \phi/\omega_i$ for $\phi > 0$ and ω_i known weight quantity. The mean and variance of y_i are expressed via the equation (1). Let $L_i = logf(y_i; \theta_i, \phi)$ and $L = \sum_i L_i$.

$$L_i = (y_i\theta_i - b(\theta_i))/\alpha(\phi) + c(y_i, \phi) \Rightarrow \frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\alpha(\phi)},\ \frac{\partial^2 L_i}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{\alpha(\phi)}$$

$$E\Big(\frac{\partial L}{\partial \theta}\Big) = 0 \quad and \quad -E\Big(\frac{\partial^2 L}{\partial \theta^2}\Big) = E\Big(\frac{\partial L}{\partial \theta}\Big)^2.$$

From the first formula

$$E\Big[\frac{y_i - b'(\theta_i)}{\alpha(\phi)}\Big] = 0 \Rightarrow \mu_i = E(y_i) = b'(\theta_i).$$

From the second formula

$$\frac{b''(\theta_i)}{\alpha(\phi)} = E\Bigg[\frac{y_i - b'(\theta_i)}{\alpha(\phi)}\Bigg]^2 = \frac{var(y_i)}{\alpha(\phi)^2} \Rightarrow var(y_i) = b''(\theta_i)\alpha(\phi).$$

The systematic component involves a vector $(\eta_1, \ldots, \eta_n)$ that is related to the linear combination of the explanatory variables; a linear model. Each component of the vector is given by the formula

$$\eta_i = \sum_j \beta_j x_{ij}, i = 1, \ldots, N.$$

This is the linear predictor of the generalised linear model.

As for the link function, this third component of the GLM breaches the systematic component with the random component. To elaborate, let $\mu_i = E(Y_i),\ i = 1, \ldots, N$ and $\eta_i,\ i = 1, \ldots, N$ be the linear predictor, the GLM links μ_i to η_i by $g(\mu_i) = \eta_i$, where $g()$ is the link function; monotonic and differentiable. Therefore,

$$g(\mu_i) = \sum_j \beta_j x_{ij},\ i = 1, \ldots, N.$$

When $g(\mu) = \mu$ the link function is called the identity link function and it leads in the linear regression model with normally distributed response variable. Furthermore, the link function that turns the mean to the natural parameter θ is called the canonical link

$$Q(\theta_i) = g(\mu_i) = \sum_j \beta_j x_{ij}.$$

1.1.1 The Likelihood function of the GLM

In order to obtain maximum likelihood estimates for the model parameters it is necessary to derive general expressions for the likelihood function of the GLM. For N independent observations the log likelihood is given by

$$L(\boldsymbol{\beta}) = \sum_i L_i = \sum_i logf(y_i; \theta_i, \phi) = \sum_i \frac{y_i\theta_i - b(\theta_i))}{\alpha(\phi)} + \sum_i c(y_i; \phi). \quad (1.3)$$

The parameter vector β comprises the model parameters of the GLM.

So, $\eta_i = \sum_j \beta_j x_{ij} = g(\mu_i)$ with link function g and likelihood equations that are given by

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} \frac{\partial L_i}{\partial \beta_j} = 0, \quad for\ all\ j.$$

The log likelihood is differentiated using the chain rule of differentiation

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{1.4}$$

On the grounds that, $\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - \mu_i}{\alpha(\phi)}\frac{\alpha(\phi)}{var(y_i}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{var(y_i}\frac{\partial \mu_i}{\partial \eta_i}$.

The Likelihood equations over the N observations for a GLM are given by the formula:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \dots, N. \tag{1.5}$$

As for the matrix form of the likelihood equations of the GLM, let V denote a diagonal matrix of variances and D denote a diagonal matrix with elements the partial derivatives $\frac{\partial \mu i}{\partial \eta_i}$. For the generalised linear model with $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ with a model matrix $\boldsymbol{X}$, these are the likelihood equations:

$$\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}. \tag{1.6}$$

1.1.2 Normal Distribution of the GLM Parameter Estimators

For large samples, the maximum likelihood estimator show a foundational property, under standard regulatory conditions; the Maximum Likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is efficient and approximately normally distributed. It is necessary to calculate the covariance matrix of that distribution which is derived from the inverse of the information matrix J.

$$E\bigg(-\frac{\partial^2 L_i}{\partial \beta_i \partial \beta_j}\bigg) = E\bigg[\bigg(\frac{\partial L_i}{\partial \beta_i}\frac{\partial L_i}{\partial \beta_j}\bigg)\bigg].$$

Substituting the equation (4) the result is

$$E\bigg(-\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\bigg) = E\bigg[\frac{(y_i - \mu_i)x_{ih}}{var(y_i)}\frac{\partial \mu_i}{\partial \eta_i}\frac{(y_i - \mu_i)x_{ij}}{var(y_i)}\frac{\partial \mu_i}{\partial \eta_i}\bigg] = \frac{x_{ij}x_{ih}}{var(y_i)}\bigg(\frac{\partial \mu_i}{\partial \eta_i}\bigg)^2$$

$$\Rightarrow E\bigg(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_j}\bigg) = \sum_{i=1}^{N}\frac{x_{ij}x_{ih}}{var(y_i)}\bigg(\frac{\partial \mu_i}{\partial \eta_i}\bigg)^2.$$

Let W be a diagonal matrix with elements

$$w_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{var(y_i)}.$$

Then the information matrix J can be transformed, with the model matrix X, to:

$$\boldsymbol{J} = \boldsymbol{X^T W X}. \tag{1.7}$$

The link function g impacts directly the form of $W\ and\ J$ as $g'(\mu_i) = \frac{\partial \mu_i}{\partial \eta_i}$ Therefore, it is concluded explicitly that

$$\beta \sim N(\boldsymbol{\beta, (X^T W X)^{-1}})\ approximately, \tag{1.8}$$

where $\boldsymbol{W}$ is a diagonal matrix with $w_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{var(y_i)}$.

1.1.3 Methods of Inference for GLM Parameters

The hypothesis test that is conducted for the coefficient of every GLM parameter is

$$H_0 : \beta = \beta_0 \quad vs \quad H_1 : \beta = \beta_0.$$

There are three different ways by which the statistical test for the significance of the model parameter (coefficient) is conducted.

Likelihood-Ratio Test

The first approach utilises the ratio of the likelihood function evaluated at β_0; l_0, over all β values permitting H_0 or H_1 to be true. The ratio create, $\Lambda = l_0/l_1$, scores values less or equal to zero due to the fact that l_0 is derived from maximisation at β_0. The likelihood-ratio test statistic is

$$-2log\Lambda = -2log(l_0/l_1) = -2(L_0 - L_1),$$

where L_0 and L_1 denote the maximised log-likelihood functions.

Under regularity conditions, it has an approximate null chi squared distribution as $N \to \infty$ with $df = 1$. The likelihood-ratio test extends to multiple parameters. To elaborate, for $\boldsymbol{\beta = (\beta_0, \beta_1)}$ the null hypothesis is $H_0 : \boldsymbol{\beta_0 = 0}$. Then l_1 is the likelihood function evaluated at $\boldsymbol{\beta}$ and l_0 is the likelihood function evaluated at $\boldsymbol{\beta}_1$ value for which the data would have been most likely when $\boldsymbol{\beta}_0 = 0$. The chi-squared test holds for the multiple parameters test with df equal to the difference in the dimensions of the parameter spaces under $H_0 \cup H_1$ and under $H_0 : \boldsymbol{\Lambda\beta = 0}$ (Agresti, 2015).

Wald Tests

The Wald Test comes next taking into account the standard errors obtained from the inverse of the information matrix. The estimated standard error is obtained by substituting with the unrestricted ML estimator of $\hat{\beta}$. The null hypothesis is $H_0 : \beta = \beta_0$ and the test statistic used is

$$z = \frac{\hat{\beta} - \beta_0}{SE},$$

which is called a Wald statistic.

Its approximate distribution is the standard normal under the null hypothesis. When it comes to multiple parameters testing $\boldsymbol{\beta = (\beta_0, \beta)}$ of the null hypothesis $H_0 : \boldsymbol{\beta_0 = 0}$ the Wald chi-squared test statistic comes below

$$\boldsymbol{\hat{\beta_0}^T [\widehat{var}(\hat{\beta_0})]^{-1} \hat{\beta_0}},$$

where $\boldsymbol{\beta_0}$ is the unrestricted ML estimate of $\boldsymbol{\beta_0}$ and $\widehat{var}(\boldsymbol{\beta_0})$ is the unrestricted estimated covariance matrix of $\boldsymbol{\hat{\beta}}$.

Score Tests

A third alternative inferential technique uses the score statistic. The score test takes advantage of the score function of the slope and the expected curvature of the log-likelihood function evaluated at β_0. The score test has the following chi-squared formulation

$$-\frac{\Big[\partial L(\beta)/\partial\beta_0\Big]^2}{E\Big[\partial^2 L(\beta)/\partial\beta_0^2\Big]} \sim \chi^2,$$

where the notation reflects derivatives with respect to β that are evaluated at β_0. From the multiparameter perspective, the score test statistic is of a quadratic form relying upon the vector of partial derivatives of the log-likelihood as well as the inverse information matrix, under H_0.

1.1.4 Deviance of the Generalised Linear Model

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the observations of the response variable and let $L(\boldsymbol{\mu}; \boldsymbol{y})$ be the log likelihood, under the GLM. For all possible models the maximum value of the log likelihood is $L(\hat{\boldsymbol{\mu}} = y; \boldsymbol{y})$ and is calculated for a model that has one parameter for each and every single observation; model is called saturated. This model is overparameterised and its predominant use is in model comparison. Due to its overfitting to the data, the saturated model is incompetent in describing the data with the least needed explanatory variables. The estimate of the mean is $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$. The deviance is given by the formula

$$-2[L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y})].$$

Comparison of Chosen Model to the Saturated

Let $\hat{\theta}_i$ be the ML estimate of each natural parameter θ_i for any chosen model, with corresponding $\hat{\mu}_i = y_i$, and $\tilde{\theta}_i$ be the estimate of each θ_i for the saturated model, with corresponding $\tilde{\mu}_i = y_i$. For maximised log-likelihoods $L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$ for the chosen model and $L(\boldsymbol{y}; \boldsymbol{y})$ for the saturated. Hence,

$$-2log\left[\frac{maximum\ likelihood\ for\ chosen\ model}{maximum\ likelihood\ for\ saturated\ model}\right] \tag{1.9}$$

is the statistic that tests whether there is strong evidence against the chosen model H_0 over a more complex H_1. Therefore it points to the lack of fit of the chosen model against the saturated-overfitted. From (1.3) it follows

$$-2[L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y}) = 2\sum_i [y_i\tilde{\theta}_i - b(\tilde{\theta}_i)]/\alpha(\phi) - 2\sum_i [y_i\hat{\theta}_i - b(\hat{\theta}_i)]/\alpha(\phi).$$

By substituting $\alpha(\phi) = \phi/\omega_i$,

$$2\sum_i \omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})/\phi.$$

$D(\boldsymbol{y};\hat{\boldsymbol{\mu}})/\phi$ is called the scaled deviance and $D(\boldsymbol{y};\hat{\boldsymbol{\mu}})$ is called deviance.

The greater the deviance the greater the lack of fit. For particular GLMs, for example the binomial and Poisson, under small dispersion asymptotics in which the number of observations is fixed and the individual observations converge to normality, the scaled deviance has a chi-squared distribution with degrees of freedom equal to the difference between the number of parameters in the saturated and the chosen model, (Agresti, 2015). The scaled deviance is preferred for model checking when ϕ is known.

Deviance Difference in LR

When $\phi = 1$, particularly in Poisson and binomial models, the deviance equals

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = -2[L(\hat{\boldsymbol{\mu}};\boldsymbol{y}) - L(\boldsymbol{y};\boldsymbol{y})].$$

Consider two nested models, M_0 with p_0 parameters and fitted values $\hat{\mu_0}$, and M_1 with p_1 parameters and fitted values $\hat{\mu_1}$ and let M_0 be a special case of M_1. The parameter space of M_0 is contained in the parameter space of M_1, (Agresti, 2015), henceforth

$$L(\hat{\boldsymbol{\mu}}_{\mathbf{0}};\boldsymbol{y}) \leq L(\hat{\boldsymbol{\mu}}_{\mathbf{1}};\boldsymbol{y}) \Rightarrow D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_{\mathbf{1}}) \leq D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_{\mathbf{0}}).$$

It is assumed that model M_0 is more preferable over M_1. The log-likelihood ratio test statistic

$$\begin{aligned}-2[L(\hat{\boldsymbol{\mu}_0};\boldsymbol{y})-L(\hat{\boldsymbol{\mu}_1};\boldsymbol{y})] &= -[L(\hat{\boldsymbol{\mu}_0};\boldsymbol{y})-L(\boldsymbol{y};\boldsymbol{y})]-\{-2[L(\hat{\boldsymbol{\mu}_1};\boldsymbol{y})-L(\boldsymbol{y};\boldsymbol{y})]\}\\ &= D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_{\mathbf{0}}) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_{\mathbf{1}})\end{aligned}$$

when $\phi = 1$.

The test statistic is large when the proposed model M_0 fits poorly the data compared to M_1.

The difference of Deviances can also be written as

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_{\mathbf{0}}) - D(\boldsymbol{y};\hat{\boldsymbol{\mu}}_{\mathbf{1}}) = 2\sum_i \omega_i[y_i(\tilde{\theta}_{1i} - \hat{\theta}_{0i}) - b(\tilde{\theta}_{1i}) + b(\hat{\theta}_{0i})].$$

1.1.5 Fitting a Generalised Linear Model

The likelihood equations (1.5) are non linear in $\hat{\boldsymbol{\beta}}$ so it is necessary to use iterative procedures to find the ML estimates.

Newton-Raphson Method

The Newton-Raphson method is used as a manner to solve non-linear equations iteratively in order to determine the value at which the function takes its maximum value. It initiates with an initial approximation of the solution. After that, it derives a second approximation by approximating the function of interest in a relatively close neighbourhood of the first approximation by a second degree polynomial and then by finding its maximum value. By repeating these steps it generates a sequence of approximations that will ultimately converge to the location of the maximum as long as the function has good geometric properties and the initial values are quite good.

Let

$$\boldsymbol{u} = \left(\frac{\partial L(\beta}{\partial\beta_1}, \frac{\partial L(\beta}{\partial\beta_2}, \dots, \frac{\partial L(\beta}{\partial\beta_p}\right)^T.$$

Let H be the Hessian matrix with elements $h_{ab} = \partial^2 L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b$. Let $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ be $\boldsymbol{u}$ and $\boldsymbol{H}$ evaluated at $\boldsymbol{\beta}^{(t)}$, approximation t for $\hat{\boldsymbol{\beta}}$. Step t in the iterative process ($t = 0, 1, \dots$) approximates $\boldsymbol{L(\beta)}$ near $\boldsymbol{\beta}^{(t)}$ by the terms up to the second order in its Taylor series expansion:

$$\boldsymbol{L(\beta) \approx L(\beta^{(t)}) + u^{(t)T}(\beta - \beta^{(t)}) + \frac{1}{2}(\beta - \beta^{(t)})^T H(t)(\beta - \beta^{(t)})}.$$

Solving

$$\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \approx \boldsymbol{u^{(t)T}(\beta - \beta^{(t)}) = 0}$$

for $\boldsymbol{\beta}$ returns the next approximation,

$$\boldsymbol{\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1}u^{(t)}} \tag{1.10}$$

for $\boldsymbol{H}^{(t)}$ being non singular. The iterations continue until the changes in the values of $\boldsymbol{L}(\beta^{(t)})$ between successive cycles are sufficiently small. The limit of $\boldsymbol{\beta^{(t)}}$ is the ML estimator. A problem with the NR method is that if the approximated function has multiple local maxima the limit is hard to find. This is the reason why a good initial approximation is essential for convergence.

Fisher Scoring Method

Fisher scoring is an alternative to Newton Raphson for solving likelihood equations. Fisher scoring uses the expected Hessian matrix or the expected information, instead of the Hessian matrix itself that is used in Newton-Raphosn (Agresti, 2015). Let $J^{(t)}$ be the sequence of approximation for the ML estimate of the expected information matrix. $J^{(t)}$ has elements $-E(\frac{\partial^2 L(\beta)}{\partial\beta_\alpha\beta_b})$, evaluated at $\boldsymbol{\beta}^{(t)}$. The Fisher scoring formula is

$$\boldsymbol{\beta^{(t+1)} = \beta^{(t)} + (J^{(t)})^{-1}u^{(t)} \Rightarrow J^{(t)}\beta^{(t+1)} = J^{(t)}\beta^{(t)} + u^{(t)}}. \tag{1.11}$$

It is easy to turn (1.11) to the matrix form $\boldsymbol{J = X^T W X}$, where $\boldsymbol{W}$ is diagonal with elements $w_i = (\frac{\partial\mu_i}{\partial\eta_i})^2/var(y_i)$. By the same token, $\boldsymbol{J^{(t)} = X^T W^{(t)} X}$, where $\boldsymbol{W}^{(t)}$ is $\boldsymbol{W}$ evaluated at $\boldsymbol{\beta}^{(t)}$. The estimated asymptotic covariance matrix J^{-1} of $\hat{\boldsymbol{\beta}}$ occurs as a by-product of the algorithm as $\boldsymbol{(J^{(t)})^{-1}}$ for the value of which convergence is adequate. For GLMs with a canonical link function, the observed and the expected information are the same; as will be shown in following section.

An easy manner to start both iterative processes takes as initial estimate of $\boldsymbol{\mu}$ the data $\boldsymbol{y}$, which is smoothed to exclude boundary values. This makes the initial estimate of the weight matrix $\boldsymbol{W}$ and subsequently the initial approximation for $\hat{\boldsymbol{\beta}}$.

Iteratively Reweighted Least Squares

There is a relation between the Fisher scoring iterative process to obtain the maximum likelihood estimations and the weighted least squares approach to estimation. The general linear model is

$$\boldsymbol{z} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

When the covariance matrix of ϵ is $\boldsymbol{V}$ the generalised least squares estimator of $\boldsymbol{\beta}$ is proved to be

$$(\boldsymbol{X^T V^{-1} X})^{-1} \boldsymbol{X^T V^{-1} z}$$

When $\boldsymbol{V}$ is diagonal this is called the weighted least squares estimator. From (1.6) the score vector for a GLM is $\boldsymbol{X^T D V^{-1}(y - \mu)}$. Take into account that $\boldsymbol{D} = diag\frac{\partial \mu_i}{\partial \eta_i}$ and $\boldsymbol{W} = diag[(\frac{\partial \mu_i}{\partial \eta_i})^2/var(y_i)]$, it turns out that $\boldsymbol{DV^{-1} = WD^{-1}}$ and the score function can be written as

$$\boldsymbol{u} = \boldsymbol{X^T W D^{-1}}(\boldsymbol{y} - \boldsymbol{\mu}).$$

Since $\boldsymbol{J} = \boldsymbol{X^T W X}$, it entails that the Fisher scoring formula

$$\boldsymbol{\beta}^{(t)} + (\boldsymbol{J}^{(t)})^{-1}\boldsymbol{u}^{(t)} = (\boldsymbol{X^T W}^{(t)}\boldsymbol{X})\boldsymbol{\beta}^{(t)} + \boldsymbol{X^T W}^{(t)}(\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$$

$$= \boldsymbol{X^T W}^{(t)}\Big[\boldsymbol{X\beta}^{(t)} + (\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)})\Big] = \boldsymbol{X^T W}^{(t)}\boldsymbol{z}^{(t)}$$

where $\boldsymbol{z}^{(t)}$ has elements

$$z_j^{(t)} = \sum_j x_{ij}\beta_j^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}$$

The Fisher scoring equations are of the form

$$(\boldsymbol{X^T W}^{(t)}\boldsymbol{X})\boldsymbol{\beta}^{(t+1)} = \boldsymbol{X^T W}^{(t)}\boldsymbol{z}^{(t)}.$$

These are the normal equations for using weighted least squares to fit a linear model for a response variable $\boldsymbol{z}^{(t)}$, when the model matrix is $\boldsymbol{X}$ and the inverse of the covariance matrix is $\boldsymbol{W}^{(t)}$. The equations have the solution

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X^T W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X^T W}^{(t)}\boldsymbol{z}^{(t)}.$$

The vector $\boldsymbol{z}^{(t)}$ in this formulation is an estimated linearised form of the link function g, evaluated at $\boldsymbol{y}$,

$$g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)})g'(\mu_i^{(t)}) = \eta_i^{(t)} + (y_i - \mu_i^{(t)})\frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = z_j^{(t)}. \quad (1.12)$$

The adjusted response variable z has each element i which is approximated by $z_i^{(t)}$ for the t cycle of the iterative process. That particular scheme regresses $\boldsymbol{z}^{(t)}$ on $\boldsymbol{X}$ with weight (i.e. inverse covariance) $\boldsymbol{W}^{(t)}$ to obtain a new approximation $\boldsymbol{\beta}^{(t+1)}$. This estimate returns a new linear predictor value $\boldsymbol{\eta}^{(t+1)} = \boldsymbol{X\beta}^{(t+1)}$ and a new approximation $\boldsymbol{z}^{(t+1)}$ which will be used to the adjusted response for the next cycle. The ML estimator stems from the repetitive use of the weighted least squares, in which the weight matrix updates at each turn. The process is called iteratively reweighted least squares (IRLS). The weight matrix $\boldsymbol{W}$ used in $\boldsymbol{var(\hat{\beta}) \approx (X^T W X)^{-1}}$, and in Fisher scoring is the inverse covariance matrix of the linearised form $\boldsymbol{z = X\beta + D^{-1}(y - \mu)}$ of $g(\boldsymbol{y})$. At convergence,

$$\boldsymbol{\hat{\beta} = (X^T \hat{W} X)^{-1} X^T \hat{W} \hat{z}}$$

for the estimated adjusted response $\boldsymbol{\hat{z} = X\hat{\beta} + \hat{D}^{-1}(y - \hat{\mu})}$.

Simplifications for Canonical Link Functions

For the GLMs which use canonical link functions there exist certain simplifications. For this,

$$\eta_i = \theta_i = \sum_{i=1}^{p} \beta_j x_{ij}$$

and

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i}{\partial \theta_i} = b''(\theta_i).$$

Since $var(y_i) = b''(\theta_i)\alpha(\phi)$,the contribution to likelihood equations from the equation

$$\frac{\partial L_i}{\partial \theta_i} = \frac{(y_i - \mu_i)x_{ij}}{var(y_i} \frac{\partial \mu_i}{\partial \eta_i}$$

for β_j simplifies to

$$\frac{\partial L_i}{\partial \beta j} = \frac{(y_i - \mu_i)}{var(y_i)} b''(\theta_i) x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{var(y_i)} \tag{1.13}$$

Often $\alpha(\phi)$ is identical for all observations, to exemplify $\alpha(\phi) = 1$ for binomial and Poisson GLMs. Then the likelihood equations are

$$\sum_{i=1}^{N} x_{ij} y_i = \sum_{i=1}^{N} x_{ij} \mu_i, \quad j = 1, \ldots, p. \tag{1.14}$$

The $\sum_{i=1}^{N} x_{ij} y_i$ are the sufficient statistics for the parameters β_j, so for the GLMs with canonical link function, the likelihood equations equate the sufficient statistics for the model parameters to their expected values (Agresti, 2015).

From the expression (1.13) for $\frac{\partial Li}{\partial \beta_j}$ with the canonical link function the second partial derivatives of the log likelihood are

$$\frac{\partial^2 L_i}{\partial \beta_h \beta_j} = -\frac{x_{ij}}{\alpha(\phi)} \frac{\partial \mu_i}{\partial \beta_h}.$$

This does not dependent on y_i, so

$$\frac{\partial^2 L_i}{\partial \beta_h \beta_j} = E\left[\frac{\partial^2 L_i}{\partial \beta_h \beta_j}\right].$$

Hence, $\boldsymbol{H = -J}$, the Newton-Raphson and Fisher Scoring algorithms are identical for GLMs that use the canonical link function, Nelder and Wedderburn (1972).

1.1.6 Model Selection

GLM model selection deals with the same issues as for the ordinary linear regression. The more explanatory variables added to the model the more complex it becomes. Nonetheless, a more complex model tends to explain better the data and provide a deeper insight taking into account the possible effects and interactions that emerge the more variables are added. However, a balance must be found between using a model with too many variables; in other words an overfitted model, and modelling the data with less predictors than needed.

Most research studies are designed to answer certain questions. Those questions lead to the choice of model terms. To elaborate, in a confirmatory analysis, a study hypothesis about an effect may be tested by comparing models with and without that effect. For exploratory research, a search among possible models may provide clues about the structure of effects and raise questions for future research. It is highly recommended to study the effects of each predictor in the response variable using descriptive statistics to get a feel of the effects and then go on with statistical modelling.

Forward and Backward Variable Selection

For p explanatory variables, the number of possible models is 2^p as each variable is included or not in the model. The best selection of explanatory variables identifies the model that scores best for a chosen criterion; as, for example, for the maximisation of the adjusted R^2 value. When the number of available explanatory variables is large then the whole procedure becomes computationally cumbersome.

The approach of forward variable selection adds terms in the model successively. At each turn it opts for the term that provides the model with the best fit. There is a point at which the more variables added in the model do not provide anything to the data fit, and the value of the R^2 adjusted reduces on the grounds that the variables already included explain the new ones. The process stops once the variables added do not provide any improvement to the model fit. A stepwise alteration of this method rechecks at each level whether the added terms at previous stages are still needed.

When it comes to backward elimination, the approach initiates with a fully complex model and removes terms step by step. The variables removed are those with the least devastating effect on the model; in other words the least important for the model. The independent variables with the highest p-values are deducted first up to the point that the leftovers are the statistical significant variables.

Both approaches have different initiations but ultimately they should both yield the same result as both procedures are considered to be equivalent. Whichever is chosen, an interaction term should not be in a model without its component main effects. On further notice, for qualitative predictors with more than two categories the process should take into account the variable as whole not sporadic indicators, at any stage. The qualitative variable should be dropped as a whole not only parts of it.

For any method used, the statistical significance should not be the only criterion by which a variable is added or excluded from the model; there are still practical significance criteria based on the research theoretical background. It is of supreme importance including a variable of great interest in the particular study in the statistical model and report its estimated effect even if it may not be of statistical significance. In case the variable is a potential confounder, including it in the model may help to reduce bias in estimating relevant effects of key explanatory variables.

The bias-variance tradeoff

Any model is a simplification of reality, as a matter of fact the scientist that believes that by choosing a model he/she has identified the 'correct' one among a set of candidates is badly mistaken. The more complex models do not guarantee a better fit to the data. In fact, a simple model with adequate fit has the advantage of parsimony, including a tendency to provide more accurate estimates of the quantity of interest. It needs to be balanced how complex the chosen model should be over the variance of an estimator and its bias. The bias occurs when the true values $E(y_i)$ differ from the values μ_{Mi} corresponding to fitting model M to the population. Choosing a simple model results in an increase in the bias; the difference between the model-based means and the true means tends to be higher. However, it is downplayed by the decreased variance that stems from the decreased number of parameters.

Many models can be consistent with the data. As a matter of fact, it is logically inconsistent to select a model that best fits the data and act as if the model was set up before the data analysed. Despite of the fact that this is common practice, it underestimate uncertainty and exaggerates the significance of the model and the model variables themselves.

Akaike's Information Criterion

The Akaike information criterion (AIC) judges how close a model fit is expected to be the true model. In the population, even though a simple model is farthest from the true relationship than is a more complex model, for a sample it may tend to provide a closer fit because of the advantages of parsimony. In a set of candidate models, the best is the one that tends to have sample fit closest to the true model fit.

A measure of 'closeness' is the Kullback-Leibler divergence of a model M from the unknown true model. Let $p(\boldsymbol{y})$ denote the density of the data under the true model, and let $p_M(\boldsymbol{y};\boldsymbol{\beta_M})$ be the density under model M with parameters $\boldsymbol{\beta_M}$. For a given value of the ML estimator $\hat{\beta}_M$ of β_M and for a future sample $\boldsymbol{y}^*$ from $p()$, the Kullback-Leibler divergence between the true and fitted distribution is

$$KL[p,p_M](\hat{\beta}_M) = E\left[log\frac{p(\boldsymbol{y}^*)}{p_M(\boldsymbol{y}^*;\beta\hat{M}}\right],$$

where the expectation is taken relative to the true distribution $p()$. The objective of AIC is to choose the model that minimises $E[KL[p, p_M(\hat{\beta}_M)]]$ for a set of potential models, where this expectation is taken relative to $p()$, with $\hat{\beta}_M$ as the random variable. To do this it is sufficient to minimise $E[-Elog[p_M(\boldsymbol{y}^*; \hat{\boldsymbol{\beta}}_{\boldsymbol{M}})]$ over the set of models. The true distribution to evaluate this expectation is unknown, but the expectation can be estimated. Akaike showed that when M is reasonably close to the true model, the maximised log likelihood for M is a biased estimator of $E[-Elog[p_M(\boldsymbol{y}^*; \hat{\boldsymbol{\beta}}_{\boldsymbol{M}})]$, and for large sample sizes the bias is reduced by subtracting the number of model parameters. Therefore, the optimal model minimises

$$AIC = -2[L(\hat{\beta}_M) - number\ of\ parameters\ in\ M].$$

In essence, the AIC penalises a model for having many parameters. Out of a set of candidate models, the one with the minimum of variables is identified as the optimal or the most parsimonious. The candidate models need not to be nested or even based on the same family of distributions for the random component (Agresti, 2015).

The Bayesian information criterion (BIC) comes as an alternative to AIC, by penalising more severely for the number of model parameters. It substitutes 2 by $log(n)$ as its multiple. BIC is based on a bayesian argument which of a set of models has highest posterior probability (Schwarz, 1978).

1.1.7 The Generalised Linear Model outperforms data transformation

There is wide controversy whether the GLM describe better the data than transforming data themselves. For instance, let g denote the model's link function in the paradigm of the GLM or a transformation function from the perspective of data transformation. The benefit of the GLM is that the model coefficients apply their effect directly on the $E(Y)$ after the inverse link function has been applied on it, while in the data transformation case the model coefficient effects impact the average of the transformed response $E(g(Y))$. In essence, the GLM is a linear model for the transformed mean of the response variable, which probability distribution is in the exponential family.

1.2 Models for Binary Data

For binary responses, statisticians most often assume a binomial distribution for the random component of the GLM. The binomial natural parameter is the log odds. The canonical link function for binomial data is the logit, and the respective GLM is named after it as logistic regression. The use of logistic regression in different industries has soared rapidly over the last decades. Logistic regression was firstly applied to biostudies, in particular for modelling the effects of smoking, cholesterol and blood pressure on the presence or absence of coronary disease. Social sciences followed using logistic regression for modelling opinions and behaviours.

1.2.1 Link functions for Binary Data

It is useful to distinguish between two sample size measures: a measure n_i for the number of Bernoulli trials that constitute a particular binomial observation, and a measure N for the number of binomial observations. Let $y_1, \ldots, y_N$ be independent binomial proportions, with $n_i y_i \sim bin(n_i, \pi_i)$, with y_i being the proportion of successes out of n_i independent Bernoulli trials, and $E(y_i) = \pi_i$ not depending on n_i. Let $\boldsymbol{n} = (n_1, \ldots, n_N)$ denote the binomial sample sizes. The overall number of observations is $n = \sum_{i=1}^{N} n_i$.

Binary Response; Grouped or Ungrouped

For binary data the outcome is 0 *or* 1 so the data format for the ungrouped data becomes a vector with elements 0 *or* 1 depending on the outcome. As $N \to \infty$ large-sample methods for statistical inference apply.

For grouped data, each observation is valued the same for the explanatory variable. n_i refers to the number of observations at the i setting of the explanatory variable with $i = 1, \ldots, N$. For grouped data, the number N of combinations of the categorical predictors is fixed, and large sample statistical inference methods apply as $n_i \to \infty$. The grouped data are quite useful for checking model fit, while ungrouped data can easily be turned to grouped for subjects that share the same values for explanatory variables.

The Latent Variable Threshold Model

When it comes to the Latent Variable Threshold model, ungrouped data are utilised in order to be studied. It is assumed that y_i^* is an unobserved continuous response for subject i such as $y_i^* = \sum_j \beta_j x_{ij} + \epsilon_i$, where ϵ_i are independent from a distribution with mean 0 and cdf F, and there is a threshold τ such that

$$y_i^* = 0 \; if \; y_i^* \leq \tau \; and \; \; y_i^* = 1 \; if \; y_i^* > \tau.$$

Then

$$\begin{aligned} P(y_i = 1) = P(y_i^* > \tau) = P(\sum_j \beta_j x_{ij} + \epsilon_i > \tau) = \\ 1 - P(\epsilon_i \leq \tau - \sum_j \beta_j x_{ij}) = 1 - F(\tau - \sum_j \beta_j x_{ij}). \end{aligned} \tag{1.15}$$

Without loss of generality let $\tau = 0$, then

$$\begin{aligned} P(y_i = 1) = F(\sum_j \beta_j x_{ij}), \; and \\ F^{-1}[P(y_i = 1)] = \sum_j \beta_j x_{ij}. \end{aligned} \tag{1.16}$$

Therefore, models for binary data naturally take the link function to be the inverse of the standard cdf for a family of continuous distributions for a latent variable (Agresti, 2015).

Probit, Logistic and Linear Probability Models

When F is the standard normal cdf, the link function F^{-1} is called the probit link and the GLM; as in (1.16) is called the probit model.

A logistic regression model has a link function of the form

$$F(z) = \frac{e^z}{1+e^z}. \tag{1.17}$$

The logisitc distribution is well shaped like the standard normal and is defined over the real line, with F^{-1} being its link function.

When an identity link function is applied; F^{-1} is a uniform cdf. The model for the binomial parameter π_i for observation i is,

$$\pi_i = \sum_j \beta_j x_{ij}.$$

This is called the linear probability model. This model must have linear predictor that falls between 0 and 1 so as to generate probability values in the range $[0,1]$. On top of that, an S-shaped curve for which π_i very gradually approaches 0 and 1 is more plausible. That is the reason why linear probability models are not widely used.

1.2.2 Logistic Regression

Logistic Regression has been embraced from the social sciences research due to its competency in modelling opinions and mindsets in financial statistics, as it is extremely beneficial in making assumptions in the context of credit scoring. Credit scoring deals with determining whether an individual has high probability of paying his/her bill on time given his/her annual income, the number of past overdue bills, and the debt liabilities.

In this section, properties and interpretation for the model parameters of logistic regression are presented. Given the logistic regression model,

$$\begin{aligned} \pi_i &= \frac{exp\Big(\sum_j \beta_j x_{ij}\Big)}{1+exp\Big(\sum_j \beta_j x_{ij}\Big)} \\ logit(\pi_i) &= log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_j \beta_j x_{ij}. \end{aligned} \tag{1.18}$$

For a single quantitative x with positive coefficient, the curve of the π_i has the shape of the cdf of a logistic distribution. As x_i changes π_i approaches 1 at the same rate it approaches 0 , due to logistic density symmetry. With multiple explanatory variables π_i is monotone in each explanatory variable according to its coefficient sign, because of $1-\pi_i = \left[1+exp\Big(\sum_j \beta_j x_{ij}\Big)\right]^{-1}$. The absolute value of its coefficient determines the rate of climb. It is of supreme significance to determine the magnitude of β. For a quantitative explanatory variable, the tangent to the curve at that particular point is drawn to describe the instantaneous rate of change in π_i at that point. That is,

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{exp\Big(\sum_j \beta_j x_{ij}\Big)}{1 + exp\Big(\sum_j \beta_j x_{ij}\Big)} = \beta_j \pi_i (1 - \pi_i).$$

The models slope is steepest at the point for which $\pi_i = 1/2$ and the slope decreases near 0 and 1. It is of interest to interpret the β_j for a quantitative variable. The model $logit(\pi_i) = \beta_0 + \beta_1 x_i$ is for a single explanatory variable. Then,

$$logit[P(y = 1|x = 1) - logit[P(y = 1|x = 0) = [\beta_0 + \beta_1(1)] - [\beta_0 + \beta_1(0)] = \beta_1.$$

Then e^{β_1} is the odds ratio

$$e^{\beta_1} = \frac{\frac{P(y=1|x=1)}{\left[1-P(y=1|x=1)\right]}}{\frac{P(y=1|x=0)}{\left[1-P(y=1|x=0)\right]}}.$$

By exponentiating both sides of the equation, in the multiple variables, the odds $\frac{\pi_i}{1-\pi_i}$ turn to be an exponential function of x_j. The odds multiply by e^{β_j} per unit increase in x_j, adjusting for the rest explanatory variables.

Let Y be a binary response variable and an explanatory variable X, and the probability of the response variable turning true over different values of the explanatory variable $\pi(x)$ be

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x).$$

The logistic regression that models the probability $\pi(x)$ is

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}$$

which is equivalent to the transformation

$$logit[\pi(x)] = log\frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x.$$

1.2.3 Normal Explanatory Variables lead to Logistic Model

Despite of the sampling mechanism, let the explanatory variables come from a normal distribution and be continuous for each response outcome. Given y suppose $\boldsymbol{x}$ has a $N(\boldsymbol{\mu}_y, \boldsymbol{V})$ distribution with $y = 0, 1$. By applying Bayes Theorem, $P(y = 1|\boldsymbol{x})$ yields $\boldsymbol{\beta = V^{-1}(\mu_1 - \mu_0)}$.

1.2.4 Inference for Logistic Regression

These are the likelihood equations for a GLM

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0,\ j = 1, \dots, p \tag{1.19}$$

since $var(y_i) = \pi_i(1-\pi_i)/n_i$ is the binomial proportion, the likelihood equations become

$$\sum_{i=1}^{N} \frac{(y_i-\mu_i)x_{ij}}{var(y_i)} f(\eta_i) = 0,\ j = 1,\ldots,p \tag{1.20}$$

As far as β is concerned,

$$\sum_{i=1}^{N} \frac{n_i[y_i - F(\sum_j \beta_j x_{ij})]x_{ij} f(\sum_j \beta_j x_{ij})}{F(\sum_j \beta_j x_{ij}(1-F(\sum_j \beta_j x_{ij})) - 0},\ j = 1,\ldots,p. \tag{1.21}$$

1.2.5 Likelihood Equations for Logistic GLM

The binary data with

$$F(z) = \frac{e^z}{1+e^z},\ f(z) = \frac{e^z}{(1+e^z)^2} = F(z)[1-F(z)] \tag{1.22}$$

have the likelihood equations

$$\sum_{i=1}^{N} n_i(y_i - \pi_i)x_{ij} = 0,\ j = 1,\ldots,p. \tag{1.23}$$

Let $\boldsymbol{X}$ denote the design matrix and $\boldsymbol{s}$ denote the binomial vector of 'success' totals with elements $s_i = n_i y_i$. The matrix form of the likelihood equations is

$$\boldsymbol{X^T s} = \boldsymbol{X^T E(s)}. \tag{1.24}$$

1.2.6 Logistic Regression Parameters; Covariance Matrix

The ML estimator $\hat{\boldsymbol{\beta}}$ has an asymptotic normal distribution with covariance matrix to be equal of the inverse information matrix. The information matrix of a GLM is $\boldsymbol{J} = \boldsymbol{X^T W X}$ from (1.7), where $\boldsymbol{W}$ is the diagonal matrix with elements

$$w_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{var(y_i)}.$$

When it comes to binomial observations, $\mu_i = \pi_i$ and $var(y_i) = \pi_i(1-\pi_i)/n_i$, the logistic model has $\eta_i = log[\pi_i/(1-\pi_i)]$ thus $\frac{\partial \eta_i}{\partial \pi_i} = 1/\pi_i(1-\pi_i)$. Therefore, $w_i = \eta_i \pi_i(1-\pi_i)$ and when the sample is quite large

$$\widehat{var}(\hat{\beta}) = [\boldsymbol{X^T \hat{W} X}]^{-1} = [\boldsymbol{X^T Diag}[\hat{n}_i\hat{\pi}_i(1-\hat{\pi}_i)]\boldsymbol{X}]^{-1}, \tag{1.25}$$

where $\widehat{\boldsymbol{W}} = \boldsymbol{Diag}[\hat{n}_i\hat{\pi}_i(1-\hat{\pi}_i)]$ is the $N \times N$ diagonal matrix with diagonal elements $\hat{n}_i\hat{\pi}_i(1-\hat{\pi}_i)$. For ungrouped data large sample stands for a large number of Bernoulli trials; N and for grouped data it demands large $n = \sum_i n_i$ in each case with fixed value of p. The estimated standard errors of $\hat{\boldsymbol{\beta}}$ is calculated as the squared root of the main diagonal elements of the equation (1.25).

1.2.7 Wald Approach is not Optimal

When it comes to the inference of logistic regression parameters, the Wald; likelihood ratio test or score method can be handy. To test $H_0 : \beta_j = 0$ the Wald chi-squared test(df=1) uses $(\hat{\beta}_j/SE_j)^2$, whereas the likelihood-ratio chi-squared test uses the difference between the deviances of the simpler and the complex model.

Both methods, yield similar results for large samples. Nonetheless, the Wald method has two drawbacks. First and foremost, its outcomes depend on the model parameterisation. Consider the null hypothesis $H_0 : \beta_0 = 0$ $(i.e.\pi = 0)$ when ny has a $bin(n, \pi)$ distribution with the null model be $logit(\pi) = \beta_0$. The Wald chi-squared test statistic, which uses the ML estimate of the asymptotic variance is $(\beta_0/SE)^2 = [logit(y)]^2[ny(1-y)]$. The Wald test statistic is $(y-0.5)^2/[y(1-y)/n]$. Evaluating both at logit and proportion scale it is concluded that the logit scale statistic is too conservative while the proportion is too liberal. A second pitfall is that when a true effect of the response variable of a logistic GLM is quite large, then the Wald test is less stable.

1.2.8 Model Fitting; Newton-Raphson and Fisher-Scoring

The ML equations are solved by using iterative methods; Newton-Raphon and Fisher-Scoring. The Newton-Raphson method is equivalent to the Fisher-Scoring approach on the grounds that the logit link is the canonical link. From the equations (1.4) and the inverse of (1.25), with respect to binomial 'success' counts $s_i = n_i y_i$, let

$$u_j^{(t)} = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}(\beta^{(t)}) = \sum_i (s_i - n_i \pi_i^{(t)} x_{ij}) \qquad (1.26)$$

$$h_{ab}^{(t)} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_a \partial \beta_b}(\beta^{(t)}) = -\sum_i x_{ia} x_{ib} n_i \pi_i^{(t)}(1 - \pi_i^{(t)}). \qquad (1.27)$$

Here $\boldsymbol{\pi}^{(t)}$, is derived from $\boldsymbol{\beta}_{(t)}$ through

$$\pi_i^{(t)} = \frac{exp(\sum_{i=1}^{p} \beta_j^{(t)} x_{ij})}{1 + exp(\sum_{i=1}^{p} \beta_j^{(t)} x_{ij})} \qquad (1.28)$$

From the equation (1.10) the $\boldsymbol{u}^{(t)}$ and $\boldsymbol{H}^{(t)}$ are used to obtain the next value approximation $\boldsymbol{\beta}^{(t+1)}$. Then,

$$\boldsymbol{\beta^{(t+1)} = \beta^{(t)} + \left\{X^T Diag[\eta_i \pi_i^{(t)}(1 - \pi_i^{(t)})]X\right\}^{-1} X^T(s - \mu^{(t)})}, \qquad (1.29)$$

where $\mu_i^{(t)} = n_i \pi_i^{(t)}$ and it returns $\boldsymbol{\pi}^{(t+1)}$ and so on.

Let $\boldsymbol{\beta}^{(0)}$ be an initial value, then the (1.28) returns $\pi^{(0)}$ and for $t > 0$ the iterations are calculated according to (1.29) and (1.28). Asymptotically, $\boldsymbol{\pi}^{(t)}$ and $\boldsymbol{\beta}^{(t)}$ converge to the ML estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\beta}}$. The $\boldsymbol{H^{(t)}}$ matrices converge to $\widehat{\boldsymbol{H}} = -\boldsymbol{X^T Diag}[n_i \hat{\pi}_i(1 - \hat{\pi}_i)]\boldsymbol{X}$. By (1.25), the estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is a by product of the model fit; $-\widehat{\boldsymbol{H}}^{-1}$.

From the section of the Iteratively Reweighted Least Squares, $\boldsymbol{\beta}^{(t+1)}$ has the IRLS form $\boldsymbol{(X^T V_t^{-1} X)^{-1} X^T V_t^{-1} z}$, where $z^{(t)}$ is calculated from

$$z_i^{(t)} = log\frac{\pi_i^{(t)}}{1-\pi_i^{(t)}} + \frac{s_i - n_i\pi_i^{(t)}}{n_i\pi_i^{(t)}(1-\pi_i^{(t)})}$$

where $\boldsymbol{V}_t$ is diagonal with elements $1/[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]$. The $\boldsymbol{z}^{(t)}$ is the linearised form of the link function of the logistic regression; logit, evaluated at $\boldsymbol{\pi}^{(t)}$. According to Agresti (2015) the $\boldsymbol{V}_t$, the limit of $\widehat{\boldsymbol{V}}$, has diagonal elements that estimate the variances of the asymptotic normal distributions of the sample logits for large $\{n_i\}$.

1.2.9 Goodness of Fit

The model fitted to the data; either grouped or ungrouped, needs to be tested with respect to the degree of fit. Likelihood ratio test is used to compare any proposed model with a more complex one in regards of lack of fit. A measure that can be used is the deviance statistic.

Deviance and Pearson statistics

For Generalised Linear Model the deviance is the likelihood ratio statistic which compares any proposed model to the saturated. The saturated alternative fits the data perfectly $\tilde{\pi}_i = y_i$. The likelihood ratio statistic is

$$-2log\left\{\left[\prod_{i=1}^{N}\hat{\pi}_i^{n_iy_i}(1-\hat{\pi}_i)^{n_i-n_iy_i}\right]/\left[\prod_{i=1}^{N}\tilde{\pi}_i^{n_iy_i}(1-\tilde{\pi}_i)^{n_i-n_iy_i}\right]\right\} =$$

$$2\sum_i n_iy_ilog\frac{n_iy_i}{n_i\hat{\pi}_i} + 2\sum_i(n_i-n_iy_i)log\frac{n_i-n_iy_i}{n_i-n_i\hat{\pi}_i}.$$

At the ith setting of the explanatory variables, n_iy_i is the number of successes and $n_i - n_iy_i$ is the number of failures, $i = 1,\ldots,N$.Therefore, the deviance is calculated as the sum over the $2N$ success and failure total of sums, which have the form

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum observed \times log(observed/fitted).$$

When it comes to grouped data, a Pearson statistic is necessary to be calculated that summarises the goodness of fit.

$$X^2 = \sum\frac{(observed - fitted)^2}{fitted} =$$

$$\sum_{i=1}^{N}\frac{(n_iy_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \sum_{i=1}^{N}\frac{[(n_i-n_iy_i)-(n_i-n_i\hat{\pi}_i)]^2}{n_i(1-\hat{\pi}_i)} =$$

$$\sum_{i=1}^{N}\frac{(n_iy_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} = \sum_{i=1}^{N}\frac{(y_i-\hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)/n_i}.$$

Chi Squared Test

For grouped data both the Deviance and the Pearson statistic are used as goodness of fit tests for testing the null hypothesis that the model holds. Under the null hypothesis, the tests have limiting chi squared distributions as the sample size becomes larger and larger. The number of parameters for grouped data is fixed due to the fixed number of settings N. So, the degrees of freedom for the chi-squared distribution is calculated as the difference between the number of parameters of the saturated and the chosen model; $df = N - p$. For large samples, the X^2 statistic converges to chi-squared faster than the Deviance.

The chi-squared limiting distribution is not present for ungrouped data. In fact, the deviance and the Pearson statistic can be uninformative about lack of fit (Agresti, 2015). Poor chi-squared distribution is assumed with grouped data having a large N with few observations at every setting. Nonetheless, a large value of the test statistics is an indicator of lack of fit but does not provide any information about its nature. The comparison between two models is essential on the grounds that the lack of fit is attributed to including or excluding particular predictors from the fitted model.

The deviance is not useful for testing model fit for the ungrouped. It is good to know that the difference of deviances can be used for grouped or ungrouped data. Let M_0 have p_0 parameters and the more complex model M_1 have $p_1 > p_0$ parameters. The difference of deviances is the likelihood ratio statistic for comparing the simple with the complex model. Under the null hypothesis, the difference has an approximate chi-squared distribution with degrees of freedom $df = p_1 - p_0$.

1.2.10 The Probit and Complementary Log-Log Models

There is an alternative to the link function used for logistic regression. The normal distribution and a skewed distribution can be used as an alternative to using the logistic distribution for the cdf inverted to obtain the link function.

Probit Models

When the link function of a binary response model is the inverse of the standard normal cdf Φ is called the probit model. It is

$$\Phi^{-1}(\pi_i) = \sum_{j=1}^{p} \beta_j x_{ij} \Leftrightarrow \pi_i = \Phi\left(\sum_{i=1}^{p} \beta_j x_{ij}\right).$$

The likelihood equations for a probit model replace Φ and ϕ in the general equations for GLM of binary data. The estimated large-sample covariance matrix of $\hat{\boldsymbol{\beta}}$ has the form

$$\widehat{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\boldsymbol{T}}\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$$

where $\hat{\boldsymbol{W}}$ is the diagonal matrix with estimates of $w_i = (\partial\mu_i/\partial\eta_i)^2/var(y_i)$. It is substituted $\mu_i = \pi_i = \Phi(\eta_i) = \Phi(\sum_j \beta_j x_{ij})$

$$\hat{w_i} = n_i\left[\Phi\left(\sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)\right]\Bigg/\left\{\Phi\left(\sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)\left[1-\Phi\left(\sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)\right]\right\}.$$

The equations are solved with the use of iterative methods of Fisher Scoring and Newton-Raphson. Both methods yield the ML estimates, nonetheless the Newton-Raphson returns slightly different standard errors on the grounds that it inverts the observed information matrix as to approximate the covariance matrix in contrast to Fisher Scoring that takes into account the expected information. Once the link functions are anything but the canonical link the former methods differ.

Log-Log and Complementary Log-Log Models

The logit and probit links are symmetric around 0.50 on the grounds that

$$link(\pi_i) = -link(1-\pi_i).$$

We have that

$$logit(\pi_i) = log\frac{\pi_i}{1-\pi_i} = -log\frac{1-\pi_i}{\pi_i} = -logit(1-\pi_i).$$

This means that the response curve for π_i is symmetric around $\pi_i = 0.5$. Both logit and probit models are inappropriate when this does not hold. The shape of the response curve is given by the model,

$$\pi_i = 1 - exp\left[-exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)\right]. \tag{1.30}$$

In that particular model the response curve is asymmetric; π_i is approaching 0 slowly but it is approaching 1 quite fast. For that model,

$$log[-log(1-\pi_i)] = \sum_{j=1}^{p}\beta_j x_{ij}.$$

This model's link function is called the complementary log-log link, due to the fact that the log-log link applies to the complement of the π_i.

A relative model to (1.30) is

$$\pi_i = exp\left[-exp\left(-\sum_{j=1}^{p}\beta_j x_{ij}\right)\right]$$

The log-log link function is applied to the GLM

$$-log[-log(\pi_i)] = \sum_{j=1}^{p}\beta_j x_{ij}.$$

Then, π_i approaches 0 quite fast but it approaches 1 quite slowly. As a matter of fact, when the log-log model holds for the probability of success, the complementary log-log model is applied on the probability of failure but with the opposite sign of the model coefficients.

The log-log link function is a special case of the inverse cdf link that uses the cdf of Type I extreme-value distribution; known as the Gumbel distribution. The cumulative distribution function is

$$F(x) = exp\{-exp\{-(x-a)/b\}\} \quad for\ b > 0,\ -\infty < a < \infty.$$

The mode of this distribution is a, the mean is $a + 0.577b$ and the standard deviation is $1.283b$. The distribution is highly skewed to the right.The asymptotic distribution of the maximum of a sequence of independent and identically distributed continuous variables is epitomised to the term extreme value.

Fisher Scoring can turn out to be quite effective in fitting GLMs for binary data with log-log link function. Consider the model in (1.30) with a single explanatory variable x. As x increases the curve is monotone increasing for $b > 0$. The complement probability at $x + 1$ equals the complement probability at x raised to the exponentiated coefficient power; $exp(\beta)$.

A way to tell between different models of different link functions is the value of the Akaike's information criterion for the model. The one that scores the lowest is chosen among the others.

1.3 Count Data Models

A plethora of variables have counts as their potential outcome. For example, the number of times a person logs into a website and the number of visits a bank customer have made to his/hers local bank are cases of count data. The Poisson distribution is commonly assumed to model the distribution of count data. The link function to connect the systematic component with the mean is the log, thus the linear model created as a result of its application is called the loglinear model. It is feasible to adjust the model for rate when the count is based on a particular index, particularly space or time.

1.3.1 Poisson GLMs

The most common distribution assumed to model count data is the Poisson distribution on the grounds that it places its mass on the set of non negative integers. The parameter of the distribution defines the mean and variance of the data.

When y_i has a Poisson distribution, the probability mass function is

$$\begin{aligned} f(y_i;\mu_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = exp[y_i log\mu_i - \mu_i - log(y_i!)] \\ &= exp[y_i\theta_i - exp(\theta_i) - log(y_i!)], \quad y_i = 0, 1, \ldots. \end{aligned} \tag{1.31}$$

It can be seen that the Poisson distribution belongs to the exponential family of distributions with $E(y) = var(y) = \mu$, which is unimodal with the mode being equal to the integer part of μ. As for its skewness it is calculated by $E(y-\mu)^3/\sigma^3 = 1/\sqrt{\mu}$, when the mean increases the Poisson distribution is less skewed and approaches the normal distribution.

The Poisson distribution is widely used for counts of events that happen at random through time or space at a certain rate. The Poisson distribution is an approximation for the binomial distribution when the number of trials is large and the probability of success is quite small. So, if $n \to \infty$ and $pi \to 0$ such that $n\pi = \mu$ is fixed, then the binomial distribution converges to Poisson.

1.3.2 Variance Stabilisation

Let $y_1, \ldots, y_n$ denote independent observations from the Poisson distribution, with $\mu_i = E(y_i)$. It is useful to transform the count data so that the variance is constant and the ordinary least squares method can be applied. Applying the delta method $g(y) - g(\mu) \approx (y-\mu)g'(\mu)$ implies that $var(g(y)) \approx [g'(\mu)]^2 var(y)$. If the response variable has a Poisson distribution then $\sqrt{y}$ has a variance of the form

$$var(\sqrt{y}) \approx \left(\frac{1}{2\sqrt{\mu}}\right)^2, \quad \mu = 4.$$

The variance approximation is more stable when the mean becomes large, in which case $\sqrt{y}$ tends to be linear in a neighbourhood around the mean μ.

On the grounds that the $\sqrt{y}$ has approximately constant variance, modelling $\sqrt{y}_i, \quad i = 1, \ldots, n$, can be easily performed using the linear regression models. Thus, the model will be linear for the $E(\sqrt{y}_i)$, not for $E(y_i)$ or $E(log(y_i))$. Using a GLM will model the mean of the data using the appropriate link function instead of modelling the mean of the function of the response.

1.3.3 Loglinear Models

The Generalised Linear Model implemented for Poisson response data is presented. The likelihood equations (1.5) for $var(y_i) = \mu_i$ and n independent observations lead a Poisson response with linear predictor $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$ and link function g to

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) = 0.$$

The GLM for count response takes as link function the logarithm. Thus, the log of the mean can be any real value. From (1.31) it is profound that the log mean is the natural parameter for the Poisson distribution hence the log link is the canonical link for a Poisson GLM, at the same time.

The loglinear model is

$$log\mu_i = \sum_{j=1}^{p} \beta_j x_{ij} \Rightarrow log\boldsymbol{\mu} = \boldsymbol{X\beta}$$

in the analytic form and the model matrix form, respectively. It is $\eta_i = log\mu_i$ and $\partial\mu_i/\partial\eta_i = \mu_i$ therefore the likelihood equations are

$$\sum_i (y_i - \mu_i)x_{ij} = 0 \tag{1.32}$$

For the Poisson GLM, the mean is given by

$$\mu_i = exp\left(\sum_{i=1}^{p} \beta_j x_{ij}\right) = (e^{\beta_1})^{x_{i1}} \dots (e^{\beta_p})^{x_{ip}}.$$

One unit increase in x_{ij} has a multiplicative impact of e^{β_j}; the mean increased by 1 is equal to the mean x_{ij} multiplied by e^{β_j}, adjusted for every variables.

1.3.4 Goodness of Fit

The likelihood equations do not have closed form solution in general. The log-likelihood function is concave and the iterative Newton-Raphson process; as well as the Fisher-Scoring method, return fitted values and estimates of the parameters. As shown in section (1.1.1) the estimated covariance matrix is

$$\widehat{var}\hat{\beta} = \boldsymbol{X^T\hat{W}X}^{-1},$$

where $\boldsymbol{W}$ is a diagonal matrix with elements $w_i = (\partial\mu_i/\partial\eta_i)^2/var(y_i) = \mu_i$. For Poisson GLMs

$$\hat{\theta}_i = log\hat{\mu}_i, \quad b(\hat{\theta}_i) = exp(\hat{\theta}_i) = \hat{\mu}_i.$$

For the saturated model,

$$\tilde{\theta}_i = logy_i, \quad b(\tilde{\theta}_i) = y_i, \quad a(\phi) = 1.$$

The Deviance of a Poisson GLM is

$$D(\boldsymbol{y}, \boldsymbol{\hat{\mu}}) = \boldsymbol{2\sum_{i=1}^{n}\left[y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i\right]}. \tag{1.33}$$

When a Poisson model includes intercept, its likelihood equation entails $\sum_i \hat{\mu}_i = \sum_i y_i$ hence $D(\boldsymbol{y}, \boldsymbol{\hat{\mu}}) = 2\sum_i [y_i log(y_i/\hat{\mu}_i)]$. Its Pearson statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

The aforementioned statistics can be used to test the goodness of fit of the Poisson GLM of interest. When the number of Poisson observations,n, is fixed and their means increase constantly the asymptotic chi-squared distributions can be assumed as distributions for the test statistics. The best approach to estimating the model that fits best the data is to compare a proposed model with a more complex and examine the extent of lack of fit.

1.3.5 Modelling Rates

In some cases, the expected value of a response count is proportional to t_i. To exemplify, the index t_i could be either the size of a population or the amount of time; in modeling crime rates for different cities, or a spatial area in modeling counts of plant species or the spread of an infectious disease. Then the sample rate is y_i/t_i with expected value μ_i/t_i. The loglinear model for the expected rate is of the form

$$log(\mu_i/t_i) = \sum_{i=1}^{p} \beta_j x_{ij}.$$

Taking into account that $log(\mu_i/t_i) = log\mu_i - logt_i$, the models makes the adjustment $-logt_i$ to the logarithmised mean. The term of adjustment,$-logt_i$, is called offset. When the model is fitted, the term $logt_i$ is used as an explanatory variable in the systematic part of $log\mu_i$, making its coefficient equal to 1.

The expected response count takes the form

$$\mu_i = t_i exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right).$$

The mean response is proportionate to t_i with a constant which depends on the explanatory variables.

1.3.6 Negative Binomial GLMs

Using a Poisson distribution to model count data, it is assumed that the variance equals the mean. In other words, the mean and the variance behave in the same manner which some times is not valid when looking at real data. The real data may exhibit variability exceeding that predicted by the assumed Poisson distribution, leading to a phenomenon called overdispersion.

Overdispersion

Heterogeneity is quite common in overdispersion problems. The mean varies according to different values of unobserved variables at fixed levels of the predictors. The heterogeneity introduces an overall distribution for the response with variance greater than that of the Poisson distribution. In the case that the variance equals the mean for every explanatory variable, it exceeds the mean when some are included. A serious drawback is that because the variance of the response must equal the mean, for certain mean values the variance cannot decrease as extra independent variables are included.

When normality is assumed, in particular in ordinary linear regression, the distribution has a separate parameter for the variance to describe variability averting overdispersion. Nonetheless, for binomial or Poisson data the variance is a function of the mean, thus it is quite common in count data. Consider the model for the mean has the appropriate link function and linear predictor but the variability of the response distribution is higher than that of the Poisson distribution. This results in the ML estimators of model parameters for Poisson distribution being still consistent but the standard errors being too small. This is the reason why an additional parameter to account for overdispersion is needed.

Negative Binomial: Poisson mixture

One simple way to account for overdispersion is a mixture model. For a fixed number of explanatory variables, given the mean λ and the distribution of the response being the Poisson, distribution with parameter varying according to covariates. Consider $\mu = E(\lambda)$ and

$$E(y) = E(E(y|\lambda)) = E(\lambda) = \mu$$

$$var(y) = E(var(y|\lambda)) + var(E(y|\lambda)) = E(\lambda) + var(\lambda) = \mu + var(\lambda) > \mu$$

Given λ, y has a $Poisson(\lambda)$ distribution and the parameter λ has the Gamma distribution. It is $E(\lambda) = \mu$ and $var(\lambda) = \mu^2/k$ for a shape parameter $k > 0$ so the standard deviation becomes proportional to the mean. The negative binomial distribution comes from the gamma mixture of the Poisson distributions for the response y with probability mass function

$$p(y;\mu,k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{\mu}{\mu+k}\right)^{y}\left(\frac{k}{\mu+k}\right)^{k}, \quad y = 0,1,\dots. \tag{1.34}$$

With k fixed, it is proved that it is a member of the exponential dispersion family appropriate for discrete variables with natural parameter $log[\mu/(\mu+k)]$ (Agresti, 2015).

If $\gamma = 1/k$ then

$$E(y) = \mu, \quad var(y) = \mu + \gamma\mu^2.$$

The $\gamma > 0$ is a dispersion parameter; the greater the value of γ the greater the overdispersion relative to the Poisson. When $\gamma \to 0, \quad var(y) \to \mu$ and the negative binomial distribution converges to the Poisson distribution.

The negative binomial distribution is more flexible than the Poisson. The mode of the Poisson distribution is the integer part of the mean and equals to 0 when $\mu < 1$. However, the negative binomial distribution is unimodal with the mode equal to 0 when $\gamma \geq 1$ otherwise it is the integer part of $\mu(1-\gamma)$. The mode can be 0 for any μ.

Negative Binomial GLMs

It is quite common to use the log link in negative binomial GLMs rather than the canonical link. Let the dispersion parameter γ be the same for all observations and consider it to be unknown. The coefficient of variation in the Gamma distribution is $\sqrt{var(\lambda)}/E(\lambda) = \sqrt{\gamma}$.

From (1.34) the log likelihood function with n independent observations is

$$\begin{aligned} L(\boldsymbol{\beta}, \gamma; \boldsymbol{y}) = & \sum_{i=1}^{n} \left[log\Gamma\left(y_i + \frac{1}{\gamma} \right) - log\Gamma\left(\frac{1}{\gamma} \right) - log\Gamma(y_i + 1) \right] \\ & + \sum_{i=1}^{n} \left[y_i log\left(\frac{\gamma\mu_i}{1 + \gamma\mu_i} \right) - \left(\frac{1}{\gamma} \right) log(1 + \gamma\mu_i) \right]. \end{aligned} \tag{1.35}$$

The likelihood equations derived from differentiating the log likelihood with respect to $\boldsymbol{\beta}$ are

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)} (\frac{\partial\mu_i}{\partial\eta_i}) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{(\mu_i + \gamma\mu_i)^2} \left(\frac{\partial\mu_i}{\partial\eta_i} \right) = 0, \quad j = 1, 2, \dots, p.$$

The equation

$$\frac{\partial^2 L(\boldsymbol{\beta}, \gamma; \boldsymbol{y})}{\partial\beta_j \partial\gamma} = -\sum_i \frac{(y_i - \mu_i)x_{ij}}{(\mu_i + \gamma\mu_i)^2} \left(\frac{\partial\mu_i}{\partial\eta_i} \right)$$

provides the Hessian matrix elements.

$$E(\partial^2 L / \partial\beta_j \partial\gamma) = 0, \ \ for\ each\ j\ and\ \beta\ and\ \gamma.$$

Thus, $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$ are asymptotically independent and the large sample standard error for $\hat{\beta_j}$ is the same regardless of the fact that γ is known or not.

The IRLS for Fisher Scoring is applied for the maximum likelihood model fitting. The estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\widehat{var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\boldsymbol{T}} \widehat{\boldsymbol{W}} \boldsymbol{X})^{-1}$$

where, with log link, $\boldsymbol{W}$ is the diagonal matrix with $w_i = (\partial\mu_i / \partial\eta_i)^2 / var(y_i) = \mu_i / (1 + \gamma\mu_i)$. The deviance of a negative binomial model is

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i \left[y_i log\left(\frac{y_i}{\hat{\mu_i}} \right) - \left(y_i + \frac{1}{\hat{\gamma}} \right) log\left(\frac{1 + \hat{\gamma} y_i}{1 + \hat{\gamma}\mu_i} \right) \right].$$

The negative binomial deviance is close to the Poisson GLM deviance when $\hat{\gamma}$ is close to 0.

Poisson-Negative Binomial Comparison

It is open to debate whether a Poisson GLM provides a better fit over a negative binomial GLM with the same explanatory variables. The choice is based on the output of the test of significance $H_0 : \gamma = 0$, on the grounds that the Poisson distribution is the special case of the negative binomial as $\gamma \rightarrow 0$.

Reparametarised Negative Binomial

By replacing the shape parameter of the gamma distribution with $k\mu$ the density distribution becomes

$$f(\lambda; k, \mu) = \frac{k^{k\mu}}{\Gamma(k\mu)} exp(-k\lambda)\lambda^{k\mu-1}, \quad \lambda \geq 0.$$

Hence, the mean and variance are $E(\lambda) = \mu$ and $var(\lambda) = \mu/k$, respectively. For that particular parametarisation the gamma mixture of Poisson distributions returns a negative binomial distribution with

$$E(y) = \mu \quad var(y) = \mu(1 + k)/k.$$

The variance becomes linear in μ and corresponds to an inflated Poisson variance that converges to it as $k \to \infty$.

Chapter 2

GAM Introduction

The generalisation of the GLM studied in the previous chapter is the Generalised Additive Model (GAM) with a linear predictor including a sum of smooth functions of covariates. The new more flexible model allows for the more flexible specification of the dependece between the response and the covariates, instead the model is specified in terms of the 'smooth functions'. It is now mandatory to depict the smooth functions and the degree of their smoothness.

2.1 Introduction

The generalised additive model has the structure of the form

$$g(\mu_i) = \boldsymbol{A_i \theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \tag{2.1}$$

where $\mu_i = E(Y_i)$ and $Y_i \sim EF(\mu_i, \phi)$, Y_i is a response variable, $EF(\mu_i, \phi)$ denotes the exponential family of distributions with mean μ_i and scale parameter ϕ, $\boldsymbol{A}_i$ is a row of the model matrix for the parametric model components, $\boldsymbol{\theta}$ is the parametric vector of the respective model components and the f_j are the smooth functions of the covariates.

2.2 Univariate Smoothing

To begin with, the representation and estimation of component functions of a model is introduced best by a model containing one function of one covariate

$$y_i = f(x_i) + \epsilon_i \tag{2.2}$$

where y_i is the response variable, x_i is a covariate, f is a smooth function and the ϵ_i are independent random variables distributed from $N(0, \sigma^2)$.

2.2.1 Function Representation with basis expansion

To estimate the smoothing function f requires that the function is represented in the appropriate manner that (2.2) becomes a linear model. This can be achieved by defining a basis; the space of functions of which the function f is an element. Choosing a basis is related to choosing some basis functions which will be treated as known; if $b_j(x)$ is the j^{th} such basis function, then the function f is assumed to be represented as

$$f(x) = \sum_{i=1}^{k} b_j(x)\beta_j \tag{2.3}$$

By substituting (2.3) into (2.2) returns a linear model.

First, take as a basis a polynomial one. Let the smoothing function be a 4^{th} order polynomial, in a way that the polynomial of order 4 and below space includes f. A basis for this space is $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$, $b_4(x) = x^3$ and $b_5(x) = x^4$ such that (2.3) becomes

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5$$

and (2.2) becomes the model

$$y_i = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + x_i^3\beta_4 + x_i^4\beta_5 + \epsilon_i.$$

The polynomial bases are useful when the research interest revolves around the properties of in the vicinity of a specified point, but the polynomial bases are quite problematic once examining the function over its whole domain. An attempt to approximate a non linear function depicts that the polynomial interpolation oscillates severely in places so that, in order to meet the requirements with respect to data interpolation and to have all derivatives to be continuous over the different values of x. If the requirement for derivatives continuity is relaxed and a piecewise linear interpolant is used instead, then a better approximation is obtained (Wood, 2017). It is sensible to use bases which are good at approximating known functions as to resemble unknown functions. By the same token, bases that perform quite well at interpolating exact points of a function are considered a good initial point for smoothing noisy observations of a function.

One basis for piecewise linear functions of a univariate variable is determined exclusively by the coordinates of the function's derivative discontinuities; particularly the locations at which the linear components come together. Consider the knots $\{x_j^* : j = 1, \ldots, k\}$ and assume that $x_j^* > x_{j-1}^*$. Then for $j = 2, \ldots, k-1$,

$$b_j(x) = \begin{cases} (x - x_{j-1}^*)/(x_j^* - x_{j-1}^*), & x_{j-1}^* < x \leq x_j^* \\ (x_{i+j}^* - x)/(x_{j+1}^* - x_j^*), & x_j^* < x \leq x_{j+1}^* \\ 0, & otherwise \end{cases} \tag{2.4}$$

while

$$b_1(x) = \begin{cases} (x_2^* - x)/(x_2^* - x_1^*), & x < x_2^* \\ 0, & otherwise \end{cases} \tag{2.5}$$

and

$$b_k(x) = \begin{cases} (x - x^*_{k-1})/(x^*_k - x^*_{k-1}), & x > x^*_{k-1} \\ 0, & otherwise \end{cases} \tag{2.6}$$

Therefore, $b_j(x)$ is zero everywhere, but the region between the knots next to either side of x^*_j.Firstly, $b_j(x)$ increases linearly from 0 at x^*_{j-1} to 1 at x^*_j and then decreases linearly to 0 at x^*_{j+1}. Basis functions with the property of being non zero only over some finite intervals have a compact support. The shape of the b_j justifies their categorisation as tent functions.

An alternative way of defining $b_j(x)$ is the linear interpolant of the data $\{x^*_i, \delta^j_i : i = 1, \ldots, k\}$ with $\delta^j_i = 1$ if $i = j$ and zero everywhere else. This makes it computationally easier for the determination of the basis.

Under the particular basis, the function $f(x)$ is represented as the linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where $X_{ij} = b_j(x_i)$.

An illustrative example follows to clarify the use of the piecewise linear basis. The research objective is to examine to what extent the common belief that a car engine with a larger cylinder capacity wears out faster than a smaller capacity. Data for 19 Volvo engines are shown in Figure 2.1

The model fit appears to be quite well but the degree of smoothness applied is arbitrary. This problem must be addressed so that theory for modelling with unknown functions is to be developed.

2.2.2 Control smoothness

One possibility is to proceed with the choosing of degree of smoothing by backward selection. Nonetheless. a particular approach is quite worrisome owing to the fact that a model based on $k-1$ evenly spaced knots will not be nested within a model based on k evenly spaced knots. It is sensible to begin with a grid of knots and simply drop knots successively as part of the backward selection, however the resulting uneven knot spacing leads to poor model performance. On top of that, the model fit of such models tends to be highly dependent on the location of the chosen knots.

One alternative is to keep the basis dimension fixed at a larger size than that believed necessary. The model's smoothness is controlled by the adding a 'wiggliness' penalty to the least squares. In particular, the model is fitted by minimising

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \sum_{i=1}^{k-1} \{f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1})\}^2,$$

instead of minimising

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2.$$

The summation term accounts for the wiggliness as a sum of squared second differences of the function at specific points. Once, the function f is quite wiggly, the penalty will take high values and when the function f is smooth, the penalty takes lower values. In case, the function f is a straight line then the penalty becomes zero. Hence, the penalty has a null space of functions that do not bear any penalty. Its dimension is 2 due to the fact that the basis for straight lines is 2-dimensional.

The parameter λ takes into account the extent at which smoothing will take place. It consists a trade off between smoothness of the approximated function f and the fidelity to the data. If $\lambda \to \infty$ a straight line estimate for the function is assumed, while for $\lambda = 0$ a unpenalised piecewise linear regression estimate holds.

For the tent functions basis, the coefficients of function f are the function values at the knots; $\beta_j = f(x_j^*)$. It becomes relatively easy to express the penalty as a quadratic form in the basis coefficients; $\boldsymbol{\beta X^T \beta}$.

Let

$$\begin{bmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \beta_3 - 2\beta_4 + \beta_5 \\ . \\ . \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & . & . & . \\ 0 & 1 & -2 & 1 & 0 & . & . \\ 0 & 0 & 1 & -2 & 1 & 0 & . \\ & . & . & . & . & . & . \\ & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \end{bmatrix} \tag{2.7}$$

so that writing the right hand side as $\boldsymbol{D}\beta$, by definition of $\boldsymbol{D}$ matrix the penalty is

$$\sum_{j=2}^{k-1} (\beta_{j-1} - 2\beta_j + \beta_{j+1})^2 = \boldsymbol{\beta^T D^T D \beta} = \boldsymbol{\beta^T S \beta} \tag{2.8}$$

where $\boldsymbol{S} = \boldsymbol{D^T D}$. Therefore, the penalised regression fitting problem is to minimise the quantity

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \beta^T S \beta. \tag{2.9}$$

The problem of estimating the smoothness degree of the model becomes the problem of estimating the smoothness parameter λ. It is better to consider the estimation of $\boldsymbol{\beta}$ given λ before addressing the estimation of the parameter λ. A straightforward expression for the minimiser is shown by Wood (2017) which accounts for the penalised least squares estimator of β,

$$\beta = (\boldsymbol{X^T X} + \lambda \boldsymbol{S})^{-1} \boldsymbol{X^T y}. \tag{2.10}$$

By the same token, the hat matrix $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X^T} + \lambda \boldsymbol{S})^{-1} \boldsymbol{X^T}. \tag{2.11}$$

It is $\hat{\boldsymbol{\mu}} = \boldsymbol{Ay}$. The above expressions are not the ones to use for computation, as orthogonal matrix methods provide greater stability. For computation purposes,

$$\left\| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X} \\ \sqrt{\lambda} \boldsymbol{D} \end{bmatrix} \beta \right\|^2 = ||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda \boldsymbol{\beta^T S \beta}.$$

The sum of squares term, on the left hand side is just a least squares objective for a model in which the model matrix has been augmented with $k - 2$ zeros. Provided that k is large enough so that the basis is more flexible than firstly anticipated, to represent the function $f(x)$, thus the exact choice of k and the selection of the knot locations do not influence the model fit. The choice of λ plays a supreme role in determining the model flexibility and the shape of $\hat{f}(x)$.

2.2.3 Choosing smoothing parameter

If λ is too high then the data will be over smoothed, while in case it is too low the data will be under-smoothed. Therefore, in both cases the estimate of the function $\hat{f}(x)$ will not be close enough to the true function $f(x)$. The best deal would be to choose the parameter λ in a fashion that $\hat{f}$ is the closest possible to the true function f. A criterion of selecting λ would pertain the minimisation of

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

where $\hat{f}_i = \hat{f}(x_i)$ and $f_i = f(x_i)$.

M cannot be directly used due to the fact that the function f is unknown. Instead, an estimate of $E(M) + \sigma^2$ can be derived , the expected squared error in predicting a new variable. Let $\hat{f}^{[-i]}$ be the model fitted to the whole data set but y_i, and define the cross validation score

$$V_0 = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - y_i)^2.$$

This is derived from leaving out one datum per turn, fitting the model to the remaining data and calculating the squared difference between the missing datum and its predicted value. Next, the squared differences are averaged over the rest of the data.

By substituting $y_i = f_i + \epsilon_i$,

$$\begin{aligned} V_0 &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i - \epsilon_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2 - 2(\hat{f}_i^{[-i]} - f_i)\epsilon_i + \epsilon_i^2. \end{aligned} \tag{2.12}$$

Because of $E(\epsilon_i) = 0$ coupled with ϵ_i and $\hat{f}_i^{[-i]}$ which are independent then it is

$$E(V_0) = \frac{1}{n}E\Big(\sum_{i=1}^{n}(\hat{f}_i^{[-i]} - f_i)^2\Big) + \sigma^2.$$

It is $\hat{f}_i^{[-i]} \approx \hat{f}$ and $E(V_0) \approx E(M) + \sigma^2$ with equality in the large sample limit. Therefore, selecting λ as to minimise V_0 is the best way to minimise M. On top of that, choosing λ to minimise V_0 is known as ordinary cross validation.

Going for ordinary cross validation is a reasonable manner to opt for even without a mean square error justification. If the best model is chosen based on the ability of the model to fit the data then it is obvious that the more complex models will be selected over the simpler.

It is computationally intensive to calculate V_0 by leaving one datum out each time, refitting the model to each of the $n-1$ observations resulting data sets. It can easily be shown that

$$V_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_i)^2/(1-A_{ii})^2,$$

where $\hat{f}$ is the estimate from fitting to all the data, and A is the influence matrix. In essence A_{ii} is often replaced by the mean, $\frac{tr(A)}{n}$, returning the generalised cross validation score

$$V_0 = \frac{n\sum_{i=1}^{n}(y_i - \hat{f}_i)^2}{(n-tr(A))^2}$$

It can be shown that the computation of the GCV has a competitive edge over the computation of the OCV, as shown in (Wahba, 1990, p.53 or sections 6.2.2 and 6.2.3, p.258).

2.2.4 The Bayesian/mixed model approach

The smoothing penalties introduction is attributed to the assumption that the truth is more likely to be smooth than wiggly. The formalisation of this belief in a Bayesian way and the specification of the prior distribution on the function wiggliness is what will be discussed in this section. The simplest choice to make is an exponential prior

$$\propto exp(-\lambda\boldsymbol{\beta^T S\beta}/\sigma^2)$$

which is recognisable as being equivalent to an improper multivariate normal prior $\boldsymbol{\beta} \sim N(\mathbf{0},\sigma^2\boldsymbol{S}^-/\lambda)$. The prior precision matrix is proportional to $\boldsymbol{S}$ on the grounds that $\boldsymbol{S}$ is rank deficient by the dimension of the penalty null space, the prior covariance matrix is proportional to the pseudo-inverse $\boldsymbol{S}^-$; $\boldsymbol{S}^-$ is defined in a way that let $\boldsymbol{S = U\Lambda U^T}$ and Λ is the diagonal matrix of the inverse of the non-zero eigenvalues with zeros in place of the inverse for any zero eigenvalues then $\boldsymbol{S^- = U\Lambda^{-1}U^T}$.

The Bayesian interpretation of the smoothing penalty provides the model with the structure of a linear mixed model and as a consequence the MAP estimate of $\boldsymbol{\beta}$ is the solution to the equations (2.9) and (2.10), while

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N(\boldsymbol{\hat{\beta}}, \boldsymbol{(X^TX+\lambda S)}^{-1}\sigma^2)$$

By introducing this form of structure to the model eases the estimation of σ^2 and λ by applying the marginal likelihood estimation; REML.

For computational purposes only a model reparameterisation is applied. The model is re-written in terms of $\boldsymbol{\beta' = D_+\beta}$ where

$$\boldsymbol{D}_+ = \begin{pmatrix} \boldsymbol{I}_2 & \mathbf{0} \\ & \boldsymbol{D} \end{pmatrix}. \tag{2.13}$$

It is $\boldsymbol{X\beta} = \boldsymbol{XD_{+}^{-1}\beta'}$ and $\boldsymbol{\beta^T S \beta} = \sum_{i=3}^{k} \beta_i'^2$. If the first couple of elements of $\boldsymbol{\beta'}$ are re-written as $\boldsymbol{\beta}^*$ and the remainder as $\boldsymbol{b}$, the Bayesian smoothing prior becomes $\boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2/\lambda)$. $\boldsymbol{\beta}^*$ is unpenalised so it is treated as a vector of fixed effects. Let $\boldsymbol{X}^*$ be a matrix consisting of the first 2 columns of $\boldsymbol{XD_{+}^{-1}}$ while $\boldsymbol{Z}$ is the matrix of the remaining columns. The smooth model becomes

$$\boldsymbol{y} = \boldsymbol{X}^* \boldsymbol{\beta}^* + \boldsymbol{Zb} + \boldsymbol{\epsilon}, \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2/\lambda), \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2).$$

2.3 Additive Models

Consider two independent variables, x and u, and a response variable y. A simple additive model is

$$y_i = \alpha + f_1(x_i) + f_2(u_i) + \epsilon_i \tag{2.14}$$

appropriate where α is the intercept, the f_j are the smooth functions and the ϵ_i are independent $N(0, \sigma^2)$ random variables.

First the additive assumption effect, $f_1(x) + f_2(u)$, is a quite strong special case of the general smooth function of two variables $f(x, u)$. On top of that, the model now contains more than one function. As a result an identifiability issue is to be addressed: f_1 and f_2 are each exclusively estimable to within an additive constant. On that note, any constant could be added to f_1 and be subtracted from f_2 without changing the predictions of the model. It is foregone conclusion that identifiability constraints need to be imposed before fitting.

Once the identifiability issue is over, the additive model can be represented via penalised regression splines, estimated by penalised least squares and the degree of smoothing chosen by cross validation or (RE)ML, in the same manner as for the simple univariate model.

2.3.1 Penalised piecewise regression representation

Every smooth function in (2.13) can be easily represented by penalised piecewise linear basis. So,

$$f_1(x) = \sum_{j=1}^{k1} b_j(x)\delta_j,$$

where δ_j are unknown coefficients, while the $b_j(x)$ are basis functions of the form (2.4), defined using a sequence of knots k_1, x_j^* evenly spaced over the range of x. Similarly,

$$f_2(x) = \sum_{j=1}^{k2} B_j(u)\gamma_j,$$

where γ_j are unknown coefficients, while the $B_j(x)$ are basis functions of the form (2.4), defined using a sequence of knots k_2, u_j^* evenly spaced over the range of u.

Consider the vector $\boldsymbol{f}_1 = [f_1(x), \ldots, f_n(x)]^T$. It is $\boldsymbol{f}_1 = \boldsymbol{X}_1\boldsymbol{\delta}$, where $b_j(x_i)$ is the element i, j of $\boldsymbol{X}_1$. Similarly, $\boldsymbol{f}_2 = \boldsymbol{X}_2\gamma$, where $B_j(u_j)$ is the i, j element of $\boldsymbol{X}_2$. A penalty as in (2.8) is also intertwined with each function: $\boldsymbol{\delta}^T\boldsymbol{D}_1^T\boldsymbol{D}_1\boldsymbol{\delta} = \boldsymbol{\delta}^T\bar{\boldsymbol{S}}_\mathbf{1}\boldsymbol{\delta}$ for f_1 and $\gamma^T\boldsymbol{D}_2^T\boldsymbol{D}_2\gamma = \gamma^T\bar{\boldsymbol{S}}_2\gamma$ for f_2.

The identifiability issue has to be addressed. Almost any linear constraint that resolved the problem could be used, however most choices lead to wider confidence intervals. The best constraints from that perspective sum to zero

$$\sum_{i=1}^{n} f_1(x_i) = 0 \iff \mathbf{1}^{\boldsymbol{T}}\boldsymbol{f}_1 = 0,$$

where $\mathbf{1}$ is an n-dimensional vector of 1's. That particular constraint allows f_1 to have the exactly same shape as before the restriction is applied; with the same penalty. The only effect of the constraint is to vertically shift f_1 as for its mean value to be zero.

The application of the constraint requires that $\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}_1\boldsymbol{\delta} = 0$ for all $\boldsymbol{\delta}$, which entails $\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}_1 = 0$. The subtraction of the column mean from each column of $\boldsymbol{X}_1$ is necessary for the latter condition. Hence, a column centered matrix is defined

$$\tilde{\boldsymbol{X}}_1 = \boldsymbol{X}_1 - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}_1/n$$

and set $\tilde{\boldsymbol{f}_\mathbf{1}} = \tilde{\boldsymbol{X}}_\mathbf{1}\boldsymbol{\delta}$. The exclusive effect of the constraint is a shift in the level of $\boldsymbol{f}_1$. This is shown as:

$$\tilde{\boldsymbol{f}_\mathbf{1}} = \tilde{\boldsymbol{X}}_\mathbf{1}\boldsymbol{\delta} = \boldsymbol{X}_\mathbf{1}\boldsymbol{\delta} - \mathbf{1}\mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}_\mathbf{1}\boldsymbol{\delta}/n = \boldsymbol{X}_1\boldsymbol{\delta} - \mathbf{1}c = \boldsymbol{f}_1 - c,$$

where $c = \mathbf{1}^{\boldsymbol{T}}\boldsymbol{X}_1\boldsymbol{\delta}/n$. Thereby, the rank of the matrix $\tilde{\boldsymbol{X}}_1$ is reduced to $k_1 - 1$ elements of the k_1 vector $\boldsymbol{\delta}$ can be uniquely estimated. There needs a simple identifiability constraint that addresses this issue: a single element of $\boldsymbol{\delta}$ is set to zero, and the corresponding column of $\tilde{\boldsymbol{X}}_1$ and $\boldsymbol{D}$ is deleted. The column centred rank reduced basis will satisfy the identifiability constraint. From now on the tildes will be dropped, and the $\boldsymbol{X}_j, \boldsymbol{D}_j$ are the constrained versions.

Having defined the constrained bases for the f_j, it is easy to reshape the formula (2.13) in the manner

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X} = \mathbf{(1, \boldsymbol{X}_1, \boldsymbol{X}_2)}$ and $\boldsymbol{\beta}^T = (\alpha, \boldsymbol{\delta}^{\boldsymbol{T}}, \gamma^{\boldsymbol{T}})$. The penalties should better be expressed as a quadratic form in the coefficient vector $\boldsymbol{\beta}$, for convenience purposes. It can be easily done by padding out S_j with zeroes, as appropriate. In particular,

$$\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S}_1\boldsymbol{\beta} = (\alpha, \boldsymbol{\delta}^{\boldsymbol{T}}, \gamma^{\boldsymbol{T}}) \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{S}}_\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \alpha \\ \boldsymbol{\delta} \\ \gamma \end{bmatrix} = \boldsymbol{\delta}^{\boldsymbol{T}}\bar{\boldsymbol{S}}_\mathbf{1}\boldsymbol{\delta}$$

2.3.2 Fitting the additive model under penalised least squares

The model coefficient estimates $\hat{\beta}$ of the model (2.13) are obtained by the minimisation of the penalised least squares

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

where the smoothing parameters λ_1 and λ_2 control the weight to be given to f_1 and f_2 in order to smooth them. Consider the smoothing parameters to be given. The single smooth case would be

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X} + \lambda_1 \boldsymbol{S}_1 + \lambda_2 \boldsymbol{S}_2)^{-1} \boldsymbol{X^T y}$$

and

$$\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X^T X} + \boldsymbol{\lambda_1 S_1} + \boldsymbol{\lambda_2 S_2})^{-1} \boldsymbol{X^T}.$$

Nonetheless, these expressions are not optimal with respect to computational stability. Therefore, it is necessary to rewrite the objective as such

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \lambda_1 \boldsymbol{\beta^T S_1 \beta} + \boldsymbol{\lambda_2 \beta^T S_2 \beta} = \left\| \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{B} \end{bmatrix} \boldsymbol{\beta} \right\|^2 \tag{2.15}$$

where

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{0} & \sqrt{\lambda_1} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \sqrt{\lambda_2} \boldsymbol{D}_2 \end{bmatrix}$$

(or any matrix of the form $\boldsymbol{B^T B} = \lambda_1 \boldsymbol{S}_1 + \lambda_2 \boldsymbol{S}_2$).

In the single smooth approach, the right hand side of (2.14) is the unpenalised least squares objective for an augmented version of the model and the corresponding response data. By applying stable orthogonal matrix based methods the model can be easily fitted by a regular linear regression model.

2.4 Generalised Additive Models

The Generalised Additive Models (GAMs) follow from the additive models, in the same way as the generalised linear models stem from the linear models. The linear predictor explains some known smooth monotonic function of the expected value of the response variable, the response is free to follow any exponential family distribution, in turn, or have a known mean variance relationship allowing the deployment of the quasi-likelihood method.

Even though the additive model was estimated by penalised least squares, the GAM will be fitted by penalised likelihood maximisation; it will be accomplished in practice by penalised iterative least squares (PIRLS). For given smoothing parameters, these steps are required if convergence is to be achieved.

1. For the current linear predictor estimate $\hat{\eta}$ and corresponding estimated average response vector $\hat{\mu}$, calculate:

   $$\omega_i = \tfrac{1}{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2} \text{ and } z_i = g'(\mu_i)(y_i - \hat{\mu}_i) + \hat{\eta}_i,$$

   where

   $$Var(Y_i) = V(\mu_i)\phi \text{ and } g \text{ is the link function.}$$

2. Defining $\boldsymbol{W}$ as the diagonal matrix so that $W_{ii} = \omega_{ii}$, minimise the quantity

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

   with respect to $\boldsymbol{\beta}$ to derive the new estimate $\hat{\boldsymbol{\beta}}$. Henceforth, the updated estimates are obtained $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\beta}$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

The penalised Least Squares problem at the last step is equivalent to the problem of the simple additive model solved in (Wood, 2017, p. 148).

Given λ_1 and λ_2 it will be direct to obtain the estimate $\hat{\boldsymbol{\beta}}$, the selection of the GCV score is quite significant for the model, though. It comes naturally to opt for the GCV score for the final linear model in the PIRLS iteration. By replacing the residual sum of squares with the Pearson statistic this particular GCV score is equivalent to the usual GCV score.

Chapter 3

The Generalised Additive Model

This chapter is written with the aim of developing methods that will assist in the model building and model estimation process given the smoothing parameters. Then smoothing parameter estimation criteria are discussed along with the computational techniques that allow for the efficient estimation of these parameters, by the criteria optimisation.

3.1 Model Set Up

The equation of a generalised additive model, as discussed in the latter chapter, is of the form

$$g(\mu_i) = \boldsymbol{A}_i\boldsymbol{\gamma} + \sum_j f_j(x_{ji}), \quad y_i \sim \boldsymbol{EF}(\mu_i, \phi),$$

where $\boldsymbol{A}_i$ is the i-row of the parametric model matrix, with corresponding parameters γ, f_j is a smooth function of x_j, $EF(\mu_i, \phi)$ stands for an exponential family distribution with mean μ_i and scale parameter ϕ. The response data y_i are modelled as independent given the mean μ_i.

A generalisation of the model is

$$g(\mu_i) = \boldsymbol{A}_i\boldsymbol{\gamma} + \sum_j L_{ij} f_j(x_{ji}), \quad y_i \sim \boldsymbol{EF}(\mu_i, \phi),$$

where L_{ij} is a bounded linear function of f_j. One example could be that L_{ij} is an evaluation functional such as $L_{ij}f_j(x_j) = f_j(x_{ji})$, which returns the GAM basis. On top of that, it could be assumed that $L_{ij}f_j(x_j) = z_j f_j(x_{ji})$, where z_i is a covariate; such model terms are often known as varying coefficient terms, Hastie and Tibshirani (2017), on the grounds that $f_j(x_{ji})$ is viewed as a regression coefficient for z which varies smoothly for the different values of x_j. The 'signal regression' term constitutes an illustrative example with the corresponding term being $L_{ij}f_j(x_j) = \int f_j(x_j)k_i(x_j)dx_j$, where k_i is an observed function.

Smoothing bases and penalties are chosen for any model, therefore model matrices $\boldsymbol{X}^{[j]}$ and penalties $\boldsymbol{S}^{[j]}$ are defined; multiple penalty matrices may be defined for every f_j. If $b_{jk}(x_{ji})$ is the k-basis function for the functions f_j, then for the basic model are the elements $X_{ik}^{[j]} = b_{jk}(x_{ji})$ and $X_{ik}^{[j]} = L_{ij}b_{jk}(x_j)$ are in the general model.

An identifiability constraint needs to be applied to each smooth term that contains $\mathbf{1}$ in the span of its $\boldsymbol{X}^{[j]}$; in case the restriction is not applied then the smooth terms will be confounded with the intercept in matrix $\boldsymbol{A}$. There is an exception when it comes to smooths for which the penalty has no null space, hence $f_j \to 0$ as the corresponding smoothing parameter(s) tend to infinity. In the bottom line constraints need to be applied to all smooths in the basic model form, but not to all $L_{ij}f_j$ terms.

The identifiability restrictions of the form $\sum_i f_j(x_{ji}) = 0$ are quite usefully absorbed into the basis by reparameterisation. Let $\boldsymbol{\chi}$ and $\boldsymbol{S^*}_j$ denote the model and penalty matrix for f_j, respectively, after the reparameterisation. By putting together $\boldsymbol{A}$ and $\boldsymbol{\chi}^{[j]}$; column-wise a model matrix is created

$$\boldsymbol{X} = (\boldsymbol{A} : \boldsymbol{\chi}^{[1]} : \boldsymbol{\chi}^{[2]} : \dots).$$

The corresponding model coefficient vector, $\boldsymbol{\beta}$, includes $\boldsymbol{\gamma}$ and the individual smooth term coefficient vectors in the end. A smoothing penalty of the model is of the form

$$\sum_j \lambda_j \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S_j} \boldsymbol{\beta},$$

where λ_j is a smoothing parameter and $\boldsymbol{S}_j$ is the matrix $\boldsymbol{S}_j^*$ embedded as a diagonal block in a matrix, otherwise it contains only 0 entries, such as $\lambda_j \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S_j} \boldsymbol{\beta}$ is the penalty for f_j. As a matter of fact there may be more than one $\boldsymbol{S}_j$ for each f_j.

In that case the model has turned into an overparameterised GLM of the form

$$g(\mu_i) = \boldsymbol{X}_i \boldsymbol{\beta}, \quad y_i \sim EF(\mu_i, \phi),$$

to be estimated by the maximisation of

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_j \lambda_j \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S_j} \boldsymbol{\beta}. \tag{3.1}$$

The smoothing parameters, λ_j, control for the trade-off between the model goodness of fit and the model smoothness.

3.1.1 Estimation of $\boldsymbol{\beta}$ for given $\boldsymbol{\lambda}$

As it is showed in Wood (2017), the equation in (6.1) is immediately recognisable as the objective in the GLMM optimisation problem with $\boldsymbol{X}$ replacing $\boldsymbol{\chi}$ in the solution process. The (6.1) can be minimised under the penalised iteratively re-weighted least squares (PIRLS) iteration:

1. Initiate $\hat{\mu}_i = y_i + \delta_i$ and $\hat{\eta}_i = g(\hat{\mu})_i$, where δ_i is usually 0, but there may be a small constant reassuring that $\hat{\eta}_i$ is finite. Iterate the next two steps till convergence.

2. Compute pseudodata $z_i = g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i$ and iterative weights $w_i = \alpha(\hat{\mu}_i)/\{g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)\}$

3. Find $\hat{\boldsymbol{\beta}}$ to minimise the weighted least squares objective

$$||\boldsymbol{z} - \boldsymbol{X\beta}||^2_W + \sum_j \lambda_j \boldsymbol{\beta^T S_j \beta}$$

and update both $\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

It is defined that $||\alpha||^2_W = \boldsymbol{\alpha^T W \alpha}$ and $V(\mu)$ as the variance function by the exponential family distribution, while $\alpha(\mu_i) = [1 + (y_i - \mu_i)\{V'(\mu_i)/V(\mu_i) + g''(\mu_i)/g'(\mu_i)\}]$. An alternative approach would be to make use of Fisher scoring in which the Hessian of the log-likelihood will give its position to its expectation; setting $\alpha(\mu_i) = 1$.

3.1.2 Scale parameter estimation

The needed REML scale estimator as shown in Wood (2017, p. 251) is

$$\hat{\phi} = \frac{||\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2_W}{n - \tau}, \tag{3.2}$$

where

$$\tau = tr\{(\boldsymbol{X^T W X} + \boldsymbol{S_\lambda})^{-1}\boldsymbol{X^T W X}, \tag{3.3}$$

where $\boldsymbol{S}_\lambda = \sum_j \lambda_j \boldsymbol{S}_j$. τ is considered as the effective degrees of freedom of the model with the function $\boldsymbol{F} = (\boldsymbol{X^T W X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X^T W X}$ taking into account the weights. Given $\boldsymbol{z}$, $\boldsymbol{W, F}$ is interpreted as the mapping matrix of the unpenalised coefficient estimates to the penalised coefficient estimates so that its trace is effectively the average shrinkage undergone by the coefficients, multiplied by the coefficient numbers. The effective degrees of freedom are derived by summing the F_{ii} values corresponding to the β_i of the smooth term.

The $||\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2_W$ corresponds to the Pearson statistic such as the REML estimate $\hat{\phi}$ is the Pearson estimate of the scale parameter.

An alternative definition of the effective degrees of freedom is sometimes useful. For presentation purposes, consider the Gaussian additive model with influence matrix is $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X^T X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X^T}$ and
$\boldsymbol{F} = (\boldsymbol{X^T X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{X^T X}$.

The expected residual sum of squares for the model is

$$E(||\boldsymbol{y} - \boldsymbol{Ay}||^2) = \sigma^2\{n - 2tr(\boldsymbol{A}) + tr(\boldsymbol{AA})\} + \boldsymbol{b^T b}, \tag{3.4}$$

where $\boldsymbol{b} = \boldsymbol{\mu} - \boldsymbol{A\mu}$ represents the smoothing bias, which can be estimated as $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\mu}} - \boldsymbol{A}\hat{\boldsymbol{\mu}}$. This leads to the alternative scale/variance estimator

$$\hat{\sigma}^2 = \frac{||\boldsymbol{y} - \boldsymbol{Ay}||^2 - \hat{\boldsymbol{b}}^{\boldsymbol{T}}\hat{\boldsymbol{b}}}{n - 2tr(\boldsymbol{A}) + tr(\boldsymbol{AA})},$$

where $\tau_1 = n - 2tr(\boldsymbol{A}) + tr(\boldsymbol{AA}) = 2tr(\boldsymbol{F}) - tr(\boldsymbol{F}F)$ as the effective degrees of freedom for the model. Another approach to τ_1 is to proceed with bias correction and derive the fitted values:

$$\tilde{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{b}} = 2\hat{\boldsymbol{\mu}} - \boldsymbol{A}\hat{\boldsymbol{\mu}} = (2\boldsymbol{A} - \boldsymbol{AA})\boldsymbol{y}.$$

The bias corrected influence matrix is $2\boldsymbol{A} - \boldsymbol{AA}$, which trace is τ_1; the effective degrees of freedom for the bias corrected model.

3.2 Smoothness Selection

The estimation of coefficient $\boldsymbol{\beta}$ has been carried out conditioned on the given smoothing parameters $\boldsymbol{\lambda}$, which has to be estimated. As discussed in the previous section there are two classes of method in general use: prediction error methods, such as GCV and AIC, or marginal likelihood methods based on the Bayesian/mixed model perspective of smoothing. There are two alternative computational strategies; the smoothness selection criterion is defined and optimised for the model itself or it is defined for the working model in the PIRLS iteration procedure. That last strategy is predominantly used in the PQL procedure; does not guarantee convergence but can be particularly fast especially in big data.

3.2.1 UBRE: known scale parameter

When it comes to estimating smoothing parameters in the simple case of an additive model with constant variance, an interesting approach is to ensure that $\hat{\boldsymbol{\mu}}$ is as close to the true mean $\boldsymbol{\mu} = E(y)$. A good measure of that particular proximity is the mean square error (MSE) which is defined as:

$$M = E(||\boldsymbol{\mu} - \hat{\boldsymbol{\beta}}||^2/n) = E(||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}||^2)/n - \sigma^2 + 2tr(\boldsymbol{A})\sigma^2/n. \quad (3.5)$$

It is quite reasonable to select the smoothing parameters as to minimise M; un-biased risk estimator (Wahba, 1990, UBRE),

$$V_u(\boldsymbol{\lambda}) = ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}||^2/n - \sigma^2 + 2tr(\boldsymbol{A})\sigma^2/n, \quad (3.6)$$

which is Mallows' Cp (Mallows, 2000). The right hand side of (3.6) relies on the smoothing parameters through $\boldsymbol{A}$.

If σ^2 is known, then estimating $\boldsymbol{\lambda}$ by minimising V_u works quite well, otherwise the estimation of σ^2 becomes problematic. Substituting the approximation

$$E(||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}||^2) = \sigma^2\{n - tr(\boldsymbol{A})\}, \quad (3.7)$$

implied by (3.2), into (3.5) yields

$$M = E(||\boldsymbol{\mu} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2/n) = \frac{tr(\boldsymbol{A})}{n}\sigma^2 \quad (3.8)$$

and the MSE is $\tilde{M} = \frac{tr(\boldsymbol{A})}{n}\hat{\sigma}^2$. Consider the comparison of the one-parameter and two-parameter unpenalised models using $\hat{M}$: the two-parameter model has to reduce $\hat{\sigma}^2$ to less than half of the one parameter σ^2 estimate before it would be judged to be a refinement. Therefore, $\tilde{M}$ is not an appropriate basis as far as the model selection is concerned.

3.2.2 Cross validation: Unknown scale parameter

When the variance is unknown, the minimisation of the average square error does not work quite well. It is recommended that the smoothing parameter estimation is based on the mean square prediction error: on the average squared error in predicting a new observation y via the fitted model. The expected mean square prediction error is

$$P = \sigma^2 + M.$$

The direct dependence on σ^2 tends to mean that criteria based on P are more resistant to over-smoothing, than are criteria based on M.

By applying cross validation, the estimation of P becomes easier (Stone, 1974). By excluding a datum, y_i from the model fitting it becomes independent of the model fitted to the remaining data.The ordinary cross validation estimate of P is:

$$V_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{[-i]})^2 \tag{3.9}$$

where $\hat{\mu}_i^{[-i]}$ denotes the prediction of $E(y_i)$.

It is unnecessary to calculate V_0 by performing n model fits to obtain n terms $\hat{\mu}_i^{[-i]}$. First, take into account the penalised least squares objective which has to be minimised to derive the i-term in the OCV score:

$$\sum_{j=1,\ j\neq 1}^{n} (y_j - \hat{\mu}_j^{[-i]})^2 + Penalties.$$

Adding zero to the objective will leave the estimates that minimise it intact; the addition of the term $(\hat{\mu}_j^{[-i]} - \hat{\mu}_j^{[-i]})^2$ to derive

$$\sum_{j=1,\ j\neq 1}^{n} (y_j^* - \hat{\mu}_j^{[-i]})^2 + Penalties, \tag{3.10}$$

where $\boldsymbol{y}^* = \boldsymbol{y} - \bar{\boldsymbol{y}}^{[i]} + \bar{\boldsymbol{\mu}}^{[i]}$: $\bar{\boldsymbol{y}}^{[i]}$ and $\bar{\mu}^{[i]}$ are vectors of zeroes but the i-elements are y_i and $\hat{\mu}_i^{[-i]}$, respectively. Minimising (3.10) results in an i-prediction $\hat{\mu}_i^{[-i]}$ and also an influence matrix $\boldsymbol{A}$. So,

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

where $\hat{\mu}_i$ is from the fit to the full data. By subtracting y_i and rearranging the equation yields

$$y_i - \hat{\mu}_i^{[-i]} = (y_i - \hat{\mu}_i)/(1 - A_{ii}) \tag{3.11}$$

so the OCV becomes

$$V_0 = \frac{1}{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}. \tag{3.12}$$

The asymptotic equivalence of the Akaike's Information Criterion with the OCV is shown in Stone (1974).

Leave-several-out cross validation

This approach can be used as well, requiring a single model fit for the computations. Assume that the goal is to build a cross validation objective on the basis of leaving out subsets of data. Let α be the set of indices of one such subset in a way that $\hat{\boldsymbol{\mu}}_{\alpha}^{[-\alpha]}$ contains the predictions of the points indexed by α, when $\boldsymbol{y}_{\alpha}$ has been excluded from the model fit. Define $\boldsymbol{y}^* = \boldsymbol{y} - \bar{\boldsymbol{y}}^{[\alpha]} + \bar{\boldsymbol{\mu}}^{[\alpha]}$ then the argument is similar to the known leave-one-out method to arrive at the identity

$$\boldsymbol{y}^* = \hat{\boldsymbol{\mu}}_{\alpha}^{[-\alpha]} = (\boldsymbol{I} - \boldsymbol{A}_{\alpha\alpha})^{-1}(\boldsymbol{y}_{\alpha} - \hat{\boldsymbol{\mu}}_{\alpha})$$

from which a leave -several-out cross validation scored is computed; the matrix $\boldsymbol{A}_{\alpha\alpha}$ is the matrix consisting of rows and columns α of $\boldsymbol{A}$.

Issues with ordinary cross validation

A sensible way to estimating smoothing parameters in the OCV, but it is dominated by two potential handicaps. First and foremost, it is computationally expensive to minimise the additive model case; with multiple smoothing parameters. There is also an annoying lack of invariance (Golub et al. (1979); Wahba (1990, p. 53)).

Consider the additive model fitting objective

$$||\boldsymbol{y} - \boldsymbol{X\beta}||^2 + \sum_{i=1}^{m} \lambda_i \boldsymbol{\beta}^T \boldsymbol{S}_i \boldsymbol{\beta}.$$

Given smoothing parameters, all inferences about $\boldsymbol{\beta}$ are made based on the objective function minimisation which are identical to the inferences that would be made by using the alternative objective,

$$||\boldsymbol{Qy} - \boldsymbol{QX\beta}||^2 + \sum_{i=1}^{m} \lambda_i \boldsymbol{\beta}_i^T \boldsymbol{\beta},$$

where $\boldsymbol{Q}$ is any orthogonal matrix of appropriate dimension. Nonetheless, the two objectives give rise to different OCV scores.

3.2.3 Generalised Cross Validation

The problem with OCV emerges due to the fact that, despite parameter estimates, effective degrees of freedom and expected prediction error being invariant to the rotation of $\boldsymbol{y} - \boldsymbol{X\beta}$ by any orthogonal matrix $\boldsymbol{Q}$; the elements A_{ii} are not invariant and do not pertain in the sum (3.12). An arbitrary choice of how the model fit is performed is quite sensitive so there needs to be refined.

One technique would require to make rotations of $\boldsymbol{y} - \boldsymbol{X\beta}$ in a manner that provide a good ground for performing cross validation. It would not be recommendable to use cross validation on data in which few points have very high leverage relative to others. Thus, highly uneven A_{ii} values will make the cross validation score (3.12) to be dominated by a small proportion of the data. The selection of $\boldsymbol{Q}$ has to be done in a way that makes A_{ii} as even as possible.

It is possible to choose $\boldsymbol{Q}$ such as the A_{ii} are equal. In case $\boldsymbol{A}$ is the influence matrix for the original problem, then the rotated problem becomes

$$\boldsymbol{A}_Q = \boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{\boldsymbol{T}},$$

but if there is a matrix $\boldsymbol{B}$ such as $\boldsymbol{B}\boldsymbol{B}^{\boldsymbol{T}} = \boldsymbol{A}$ then the influence matrix becomes:

$$\boldsymbol{A}_Q = \boldsymbol{Q}\boldsymbol{B}\boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{Q}^{\boldsymbol{T}}.$$

If the orthogonal matrix $\boldsymbol{Q}$ is such that every row of $\boldsymbol{Q}B$ has the same Euclidean length, then all the elements of the diagonal matrix $\boldsymbol{A}_Q$ are equal. As a matter of fact, their value is $tr(\boldsymbol{A})/n$ because

$$tr(\boldsymbol{A}_Q) = tr(\boldsymbol{Q}\boldsymbol{A}\boldsymbol{Q}^{\boldsymbol{T}}) = tr(\boldsymbol{A}\boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{Q}) = tr(\boldsymbol{A}).$$

There must be examined whether the matrix $\boldsymbol{Q}$ exists. It is always possible to produce an orthogonal matrix to be applied from the left; its objective is to perform a rotation that affects only two rows of the targeted matrix. Such a matrix is called Givens rotation. As long as the angle of rotation; increases smoothly from zero, the Euclidean lengths of these two rows vary smoothly. However, the rotation sum of their squared lengths remains the same. Once the rotation angle reaches 90 degrees, the row lengths are interchanges which is credited to the fact that the magnitudes of the elements of the rows have been interchanged. Henceforth, there must exist a critical point at which the row lengths become equal.

3.2.4 REML and Marginal Likelihood

It is quite common to take the Bayesian approach when it comes to smoothing penalties selection as they correspond to a Gaussian prior on the model coefficients. The selection of smoothing parameters based on the maximisation of the Bayesian log marginal likelihood could be an alternative way. The log of the joint density of the data and coefficients $\boldsymbol{\beta}$ is

$$V_\tau(\boldsymbol{\lambda}) = log \int f(\boldsymbol{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})d\boldsymbol{\beta}.$$

That integral can be considered as the average likelihood of draws from the prior. Empirical Bayes is called the procedure in which the parameters estimation is performed under the condition of the log likelihood maximisation. In order to evaluate the integral a Laplace approximation is performed

$$V_\tau(\boldsymbol{\lambda}) \approx l(\hat{\boldsymbol{\beta}}) - \frac{\hat{\boldsymbol{\beta}}\boldsymbol{S}_\lambda\boldsymbol{\beta}}{2\phi} - \frac{log|\boldsymbol{S}_\lambda/\phi|}{2} - \frac{log|\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}/\phi + \boldsymbol{S}_\lambda/\phi|}{2} + \frac{M}{2}log(2\pi), \tag{3.13}$$

where l is the log likelihood, $\boldsymbol{W}$ is the diagonal matrix of full Newton weights at convergence of the PIRLS iteration, M is the dimension of the null space of S_λ and $|\boldsymbol{A}|_+$ denotes the product of non-zero eigenvalues of $\boldsymbol{A}$.

The frequentist marginal likelihood can also be used, in which the fixed effects/unpenalised components of the model are not integrated out of the likelihood. This will lead ultimately to smoother models on the grounds that the frequentist marginal likelihood tends to underestimate variance components and smoothing parameters can be thought of as precision parameters.

3.3 Estimation of smoothing parameters

The smoothing parameter estimation can be computationally intensive on the grounds that the simultaneous maintenance of efficiency and stability is quite difficult. The strategies that come into practice follow.

1. The direct optimisation of the marginal likelihood criterion. An outer iteration to optimise the smoothing parameters will be required. Every set of smoothing parameters tried by the outer iteration will require an inner PIRLS iteration to specify the model coefficient estimates for that trial. The implementation of Newton's method for maximum reliability of the outer iteration is a must.

2. Apply the Gaussian additive model version of the chosen smoothing parameter selection technique to the working penalised linear model fitted at each iteration of PIRLS of β estimation given $\boldsymbol{\lambda}$. The Central Limit Theorem is a good justification for the application of REML.

3. A simple update formula could be good; the generalised Fellner-Schall method, to update the smoothing parameters at every step of the PIRLS iteration. This optimises the Laplace approximate REML of the method

3.4 The generalised Fellner-Schall method

The generalised Fellner-Schall method is the simplest to implement; simple explicit formulae are obtained for updating the smoothing parameters in order to increase the Laplace approximate REML score of the model. The method's simplicity allows for variations of it being applied to GAM. Use of the method with smooth interaction terms was only possible using the tensor product smooths as shown in Wood (2017, p. 235). These limitations are removed by the method generalisation (Wood and Fasiolo, 2017).

The log restricted marginal likelihood of the Gaussian additive model is

$$l_r(\boldsymbol{\lambda}) = -\frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_\lambda||^2 + \hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}_\lambda}{2\sigma^2} + \frac{log|\boldsymbol{S}_\lambda/\sigma^2|}{2} - \frac{log|\boldsymbol{X}^T\boldsymbol{X}/\sigma^2 + \boldsymbol{S}_\lambda/\sigma^2|}{2} + c,$$

where $\beta_\lambda = argmax_\beta f_\lambda(\boldsymbol{y}, \boldsymbol{\beta})$for a given $\boldsymbol{\lambda}$. Given that $\partial(||\boldsymbol{y} - \boldsymbol{X}\beta||^2 + \boldsymbol{\beta}^T\boldsymbol{S}_\lambda\boldsymbol{\beta})/\partial\beta|_{\hat{\beta}_\lambda} = 0$ it is

$$\frac{\partial l_r}{\partial \lambda_j} = tr(\boldsymbol{S}_\lambda^- \boldsymbol{S}_j)/2 - tr\{(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{S}_j\}/2 - \hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_\lambda/2\sigma^2.$$

The Theorem presented in the end of the section entails that $tr(\boldsymbol{S}_\lambda^- \boldsymbol{S}_j) - tr\{\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{S}_j\}$ is non-negative, while $\hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_\lambda$ is non-negative by the positive semi-definiteness of $\boldsymbol{S}_j$. The $\frac{\partial l_r}{\partial \lambda_j}$ will be negative if

$$tr(\boldsymbol{S}_\lambda^{-1}\boldsymbol{S}_j) - tr\{(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{S}_j\} < \hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_\lambda,$$

pointing that λ_j should better be decreased. If the inequality is reversed, then $\frac{\partial l_r}{\partial \lambda_j}$ is positive indicating that λ_j should be increased. In case the inequality becomes equality then λ_j should not be modified. An update meeting the above requirements is

$$\lambda_j^* = \sigma^2 \frac{tr(\boldsymbol{S}_\lambda^- \boldsymbol{S}_j) - tr\{(\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{S}_j\}}{\hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_\lambda} \lambda_j, \tag{3.14}$$

with λ_j^* set to some predefined upper limit if $\hat{\boldsymbol{\beta}}_\lambda^T \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_\lambda$ is quite close to zero that the limit would otherwise be exceeded.

Let $\boldsymbol{B}$ be a positive definite matrix and $\boldsymbol{S}_\lambda$ be a positive semi-definite matrix parameterised by $\boldsymbol{\lambda}$ and with a null space that is independent of the value of $\boldsymbol{\lambda}$. Let positive semi-definite matrix $\boldsymbol{S}_j$ denote the derivative of $\boldsymbol{S}_\lambda$ with respect to λ_j. Then $tr(\boldsymbol{S}_\lambda^- \boldsymbol{S}_\lambda) - tr\{(\boldsymbol{B} + \boldsymbol{S}_\lambda)^{-1}\boldsymbol{S}_j\} > 0$.

3.5 Initial smoothing parameter choices

Even though the smoothing parameter estimation methods are numerically robust, it is of prime importance to choose a good starting point. The effective degrees of freedom of each smooth will be between its minimum and maximum possible values and the model coefficients will be sensitive to small scale modifications in the log smoothing parameters. Hence, the determination of the parameter update is based on the local derivative information.

This can be easily accomplished by demanding the leading diagonal elements of $\boldsymbol{X^TWX}$ and $\boldsymbol{S}_\lambda$ to be roughly balanced. If s denotes the non-zero elements of $diag(\boldsymbol{S}_j)$ and d denotes the elements of the matrix $diag(\boldsymbol{X^TWX})$; the selection of λ_j is done in a fashion that $mean(\boldsymbol{d}/(\boldsymbol{d} + \lambda_j s)) \approx 0.4$. For models that can be estimated by PIRLS, the starting estimate of W is used. In the general case, an initial estimate of the diagonal Hessian matrix of the log likelihood is replaced by $diag(\boldsymbol{X^TWX})$.

3.6 AIC and smoothing parameters

A significant part of the model selection process is considered to be the estimation of the smoothing parameters. Nonetheless, smoothing parameter selection is quite useful when it comes to comparing models that are not nested. The AIC is broadly used in model selection, but its application should take into account an additional care when it comes to containing random effects and smoothers. The two approaches follow.

1. Marginal AIC is related to the (frequentist) marginal likelihood of the model; the likelihood obtained by treating all penalised coefficients as random effects and integrating them out of the joint density of the data and random effects. Then the number of coefficients to use is the number of fixed effects and the number of variance and smoothing parameters as well.

2. Conditional AIC is backed on the likelihood of all the coefficients at their maximum penalised likelihood estimates. The number of coefficients in the penalty has to be determined from the estimate of the effective number of parameters as to take into account the coefficient penalisation.

The two approaches have a different perspective for the smooths. For the marginal likelihood the smooth is treated as a frequentist random effect; drawn from its marginal/prior distribution on each date replication. The smooths are quite often seen by statistical modellers as fixed effects, instead. That latter view is well aligned with the philosophy of conditional AIC.

An issue with the marginal AIC approach is that the frequentist marginal likelihood underestimates the variance components. The procedure is biased towards simpler models making the AIC to reflect that bias. An alternative would be to use the REML in the AIC computation but it is comparable between models with the same structure of fixed effects so that it cannot be used to compare models with or without smooths.

The conservative approach to conditional AIC which accounts for the effective degrees of freedom from section 3.1.2 in the AIC penalty term leads to an AIC extremely anti-conservative; in the random effects context the AIC can select a model which includes a random effect that is not present in the true model. Hastie and Tibshirani have shown that there is an issue with a neglect in the smoothing parameter uncertainty in τ. Numerous corrections have been proposed to deal with this issue (Greven and Kneib (2010); Yu and Yau (2012); Saefken et al. (2014)) but in this dissertation the technique of Wood et al. (2016) is followed; general enough to apply to a wide spectrum of models particularly the ones analysed later on.

The main idea is to calculate a first order correction to the posterior distribution for the model coefficients, taking into account the smoothing parameter uncertainty. Hence, the penalty term in the AIC is expressed from the Bayesian covariance matrix of the coefficients. By substituting the corrected covariance matrix into the AIC penalty returns the corrected version of τ.

3.7 Hypothesis Testing

An alternative approach to model selection is considered to be hypothesis testing, particularly when it comes to selecting simpler models in favour of more complex ones. The p-values for the parametric model effects can be easily computed in the same way as they would be for the unpenalised model. Let the null hypothesis be $H_0 : \boldsymbol{\beta}_j = 0$ where $\boldsymbol{\beta}_j$ is a subvector of $\boldsymbol{\beta}$ containing only fixed effects coefficients. The smooths are treated as random effects and the frequentist covariance matrix for $\hat{\beta}_j$ can be read from the Bayesian covariance matrix for $\boldsymbol{\beta}$ (Wood, 2017, see sections 3.4.3 and 2.4.2, p. 151 and 80). Let $\boldsymbol{V}_{\beta j}$ be the block of $\boldsymbol{V}_\beta$ corresponding to $\boldsymbol{\beta}_j$. For the more general case it is

$$\hat{\boldsymbol{\beta}}_j^T \boldsymbol{V}_{\beta_j}^{-1} \hat{\boldsymbol{\beta}}_j / p_j \sim F_{p_j, n-p},$$

if there is a scale parameter estimate involved,

$$\hat{\boldsymbol{\beta}}_j^T \boldsymbol{V}_{\beta_j}^{-1} \hat{\boldsymbol{\beta}}_j / p_j \sim \chi^2_{p_j},$$

where p and p_j are the dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_j$, respectively. The more general linear hypothesis can be tested; $H_0 : \boldsymbol{C\beta_j} = \boldsymbol{d}$ by replacing $\hat{\boldsymbol{\beta}}_j^T \boldsymbol{V}_{\beta_j}^{-1} \hat{\boldsymbol{\beta}}_j$ by $(\boldsymbol{C\hat{\beta}_j} - \boldsymbol{d})^T (\boldsymbol{C V_{\beta j} C^T})^{-1} (\boldsymbol{C\hat{\beta}_j} - \boldsymbol{d})$ in the above distributional results. For the single parameter case, the test results can be rewritten as exactly equivalent tests using t_{n-p} or $N(0,1)$ as the baseline distributions.

When it comes to the penalised terms in a model, the computation of the p-value becomes more difficult due to the effect of the penalisation term. Test statistics can be implemented for the parametric terms, but there are two main issues. The first issue is that the penalisation may affect the $\boldsymbol{V}_{\beta_j}$. In such a case a generalised inverse can be applied and the reference distribution can be modified accordingly. When the $\boldsymbol{V}_{\beta_j}$ is inverted then the components with low weight are up-weighted; the thing is that many of these components are of low variance on the grounds that they are heavily penalised and contribute almost nothing to the model. In that way the power of the test statistic is diminished. In order to develop more efficient procedures there are two main issues to consider distinctly; smooth terms where the null space of the penalty is finite dimensional and terms with no penalty null space.

3.7.1 Smooth terms and approximate p-values

Let the null hypothesis of interest be $H_0 : f_j(x) = 0$ for all x in the range of interest. The main objective is to test whether the function f_j is essentially needed in a model or not. The good interval properties rest on considering coverage across the whole interval. It is reasonable to create a test statistic backing on the whole interval.

Let $\boldsymbol{f}_j$ be the vector of $f_j(x)$ evaluated at the observed covariate values. Let $\tilde{\boldsymbol{X}}$ be such that $\boldsymbol{f}_j = \tilde{\boldsymbol{X}}\boldsymbol{\beta}$. Well calibrated confidence intervals for $\boldsymbol{f}_j$ can be obtained if the approximate result to begin with is $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{V}_\beta)$, where $\boldsymbol{V}_{\beta j}$ is the Bayesian covariance matrix for $\boldsymbol{\beta}$ so that approximately $\hat{\boldsymbol{f}}_j \sim N(\boldsymbol{f}_j, \boldsymbol{V}_{fj})$, where $\boldsymbol{V}_{fj} = \tilde{\boldsymbol{X}} \boldsymbol{V}_\beta \tilde{\boldsymbol{X}}^T$. The good calibration of the confidence interval relies on the behaviour of the smoothing bias when averaged across the function implying that a test statistic for f_j also across the function evaluation of f_j. To exemplify,

$$T_r = \hat{\boldsymbol{f}}_{\boldsymbol{j}}^T \boldsymbol{V}_{\boldsymbol{f}}^{r-}{}_j \hat{\boldsymbol{f}}_{\boldsymbol{j}},$$

where $\boldsymbol{V}_{\boldsymbol{f}}^{r-}{}_j$ is a rank r pseudo-inverse of $\boldsymbol{V}_{fj}$. An emerging issue is the selection of the rank of r. The maximum rank of $\boldsymbol{V}_{fj}$ is p_j; the number of coefficients of f_j. The rank r cannot exceed p_j, but if it is set $r = p_j$ then the power becomes poor, since the very components of f_j which are the most heavily penalised to zero are those most heavily up-weighted in the statistic, despite being the components for which the data do not carry enough information. A way to select r is to consider the number of coefficients required to best approximate the penalised estimate $\hat{f}$ by deploying an unpenalised estimate. A first order correction for the approximation of the penalised estimate is considered as straightforward on the grounds that the unpenalised estimate has no smoothing bias. Henceforth, the approximately optimal r is the τ_1 version of the term-specific distribution EDF as shown in the section 6.1.2 in Wood (2013), §2.2.

The choice of r ascertains that the most heavily penalised components of f_j are excluded from the test statistic. However, the rank might not be integer so rounding it up would be reasonable. The issue that arises in such a case is that by rounding down to the dimension of the penalty null space results in significant information loss. Imagine a weakly nonlinear $f_j(x)$ without linear trend component; modelled by a cubic regression spline. In case the effective degrees of freedom are 1.45 we would round to $r = 1$, but due to the absence of linear trend the function cannot be distinguished from zero. An easy fix would be to round up the effective degrees of freedom. To exemplify, rounding to $r = 2$ would result in a test statistic in which a nearly zero component of f_j had an utterly disproportionate influence on the test statistic which in turn leads to poor performance.

A modification of the test statistic would be essential to avoid these rounding issues; r can be set to the un-rounded EDF of f_j, while behaving exactly like the original test statistic for integer r and having similar properties whether r is integer or not. Under the null distribution, T_r follows χ^2 distribution, if r is integer; $E(T_r) = r$ and $Var(T_r) = 2r$.

Let k denote the integer part of r rounded down, $\nu = r - k$ and $\rho = \{\nu(1-\nu)/2\}^{1/2}$ and suppose that the columns of $\boldsymbol{U}$ contain the eigenvectors of $\boldsymbol{V}_{fj}$ corresponding to its non-zero eigenvalues Λ_i. The desired properties of T_r are obtained by setting

$$\boldsymbol{V}_{f_j}^{r-} = \boldsymbol{U} \begin{bmatrix} \lambda_1^{-1} & & & & \\ & \ddots & & & \\ & & \lambda_{k-2}^{-1} & & \\ & & & \boldsymbol{B} & \\ & & & & \boldsymbol{0} \end{bmatrix} \boldsymbol{U}^T, \tag{3.15}$$

where $\boldsymbol{B} = \tilde{\boldsymbol{\Lambda}} \tilde{\boldsymbol{B}} \tilde{\boldsymbol{\Lambda}}$,

$$\tilde{\boldsymbol{\Lambda}} = \begin{bmatrix} \lambda_{k-1}^{-1/2} & 0 \\ 0 & \lambda_k^{-1/2} \end{bmatrix}$$

and

$$\tilde{\boldsymbol{B}} = \begin{bmatrix} 1 & \rho \\ \rho & \nu \end{bmatrix}.$$

If $\nu_1 = \{\nu + 1 + (1 - \nu^2)^{1/2}\}/2$ and $\nu_2 = \nu + 1 - \nu_1$ (the eigenvalues of $\tilde{B}$), it is

$$T_r \sim \chi^2_{k-2} + \nu_1 \chi^2_1 + \nu_2 \chi^2_1, \tag{3.16}$$

under H_0, hence $T_r \sim \chi^2_r$ if r is integer.

The c.d.f. of the distribution in (3.16) can be calculated by the Davies method (Davies, 1980), or approximated by using a $Gamma(r/2, 2)$ distribution, which can be smoothed in the upper tail by employing the approximation of Liu et al. (2009). Once a scale parameter is estimated then the p-value is calculated as $p = pr(\chi^2_{k-2} + \nu_1 \chi^2_1 + \nu_2 \chi^2_1 > t_r \chi^2_k / k)$ where k is the residual degrees of freedom used to compute the scale estimate, and t_r is the observed T_r. An approximate distribution of T_r this can be cheaply evaluated by quadrature. The Davies method (1980) can be used to compute exactly $p = pr(\chi^2_{k-2} + \nu_1 \chi^2_1 + \nu_2 \chi^2_1 - t_r \chi^2_k / k > 0)$.

There is an ambiguity in the definition of T_r, on the grounds that $\boldsymbol{B}$ is not diagonal and eigenvectors are only unique up to a change of sign. Even though the null distributional outcome holds for either version, their actual arithmetic values are slightly different. An easy manner to overcome this problem is proposed in Wood (2017), who suggested averaging them and computing the p-values.

3.7.2 Parametric term test against a smooth alternative

It is of interest to test whether a smooth is necessary, or if the parametric terms in the null space of the smooth suffice. To elaborate, it is investigated whether a cubic spline is needed or a straight line is enough. A way to deal with it is to parameterise the smooth so that the basis for functions in the null space of the penalty is distinguished from the basis for the penalty range space. The null space terms can be included as simple parametric effects.

3.7.3 Approximated generalised likelihood ratio tests

It is quite common to try comparing GAMs using a generalised likelihood ratio test. One way is to apply the frequentist marginal likelihood taking into account the number of fixed effects in addition to the number of smoothing and variance terms as to derive the needed degrees of freedom. Another fashion is to use the likelihood coupled with the effective degrees of freedom. In case of random effects, none way yields a result which is owed to the fact that in the marginal case the null is restricting the variance parameters to the edge of the feasible parameter space, and in the conditional case the effective degrees of freedom resemble the number of unpenalised coefficients required to approximate the penalised model.

When it comes to smooth parameters, marginal likelihood tends to be biased towards slightly over-smooth estimates; inflates the acceptance rate of null models. On the contrary, the conditional manner favours the larger alternative model unless the smoothing parameter uncertainty is accounted for. The use of GLRT is justified with respect to approximating a penalised model by an unpenalised model, with a number of coefficients for each smooth given by the effective degrees of freedom for every smooth. In case that a particular approximation is considered to be a good one, the distribution of the log likelihood of each penalised model and of their difference should be quite approximated by the equivalents for the unpenalised approximations. Therefore, for comparing a null model with coefficients $\hat{\beta}_0$ to a larger model with coefficients $\hat{\beta}_1$, the following approximation can be deployed under the null hypothesis

$$\lambda = 2\{l(\hat{\beta}_1) - l(\hat{\beta}_0)\} \sim \chi^2_{EDF_1 - EDF_0}.$$

The penalised model approximation is the best possible one if the penalised model is bias corrected. Under the null model, the bias corrections of the null and alternative models should be cancelled approximately. To account for smoothing parameter uncertainty an EDF correction is deployed. It requires that REML or ML for smoothing parameter estimation.

If the GLTR is operating quite well, then the p-values have to be uniformly distributed on the unit interval. There are four computation methods deployed.

1. GCV smoothing parameter selection with the effective degrees of freedom given by τ from 6.1.2 in Wood (2017).

2. GCV smoothing parameter selection with the effective degrees of freedom given by τ_1 from 6.1.2 in Wood (2017).

3. REML smoothing parameter estimation and τ for the EDF.

4. REML smoothing parameter estimation and τ for the EDF with τ_1 corrected for smoothing parameter uncertainty, section 6.11 in Wood (2017).

As for comparing models with different random effect structure, the test fails in providing a reliable model comparison. As far as other model comparisons are concerned, the test provides a reasonable approximation given the smoothing parameter uncertainty correction is applied; the use of REML or ML smoothing parameter selection is essential.

Chapter 4

The GAMs for Location, Scale and Shape

In this chapter a general class of univariate regression models; the GAM, is developed, studied from the perspective of parameterising location, scale and shape (GAMLSS), in which case the exponential family assumption is loosened and replaced by a general distribution family. The systematic part of the model is expanded to permit the mean and the model parameters of the conditional distribution of y to be modelled as parametric and/or additive nonparametric (smooth) functions of the explanatory variables and/or random effect terms. The GAMLSS model fitting is accomplished by either one of two different algorithmic procedures. The RS algorithm relies on the algorithm that was used for the fitting of the mean and dispersion additive models as proposed in Rigby and Stasinopoulos (1996), while the second algorithm(CG) deployed is based on the Cole and Green (1992) algorithm.

4.1 The GAM for Location, Scale and Shape

4.1.1 Model Definition

The p parameters, $\boldsymbol{\theta}^T = (\theta_1, \ldots, \theta_p)$, of a population probability (density) function $f(y|\boldsymbol{\theta})$ are modelled by deploying additive models. The model assumption is, for $i = 1, \ldots, n$, observations y_i are independent conditional on $\boldsymbol{\theta}^i$, with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$, where $\boldsymbol{\theta}^{iT} = (\theta_{i1}, \ldots, \theta_{ip})$ is a vector of p parameters related to the independent variables and random effects; covariates are stochastic or the observed values depend on their past values.

Let $\boldsymbol{y}^T = (y_1, \ldots, y_n)$ be the vector of the response variable observations. On top of that, for $k = 1, 2, \ldots, p$ let $g_k(\cdot)$ be a known monotonic link function connecting $\boldsymbol{\theta}_k$ to the explanatory variables and random effects through an additive model given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{4.1}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are vectors of length n, $\boldsymbol{X}_k$ is a known design matrix of order $n \times J_k^{'}$, $\boldsymbol{Z}_{jk}$ is a fixed known $n \times q_{jk}$ design matrix and γ_{jk}; q_{jk}-dimensional random variable, for $j = 1, 2, \ldots, J_k$ combined into a single vector γ_k with a single design matrix $\boldsymbol{Z}_k$; the (4.1) is selected because it is suited to the backfitting algorithm (Rigby and Stasinopoulos, 2005, see Appendix B) and makes easy the combinations of different types of random effects terms to be included in the model.

For $k = 1, 2, \ldots, p$, $J_k = 0$ then the model (4.1) takes its reduced full-parametric form

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k. \tag{4.2}$$

If $\boldsymbol{X}_{jk} = \boldsymbol{I}_n$ and $\boldsymbol{\gamma}_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ for every combination of j and k in (4.1) this yields

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\boldsymbol{x}_{jk}), \tag{4.3}$$

where $\boldsymbol{x}_{jk}$, for $j = 1, \ldots, J_k$, and $k = 1, \ldots, p$ are vectors of length n. The h_{jk} function is an unknown function of the explanatory X_{jk} and $\boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ is the vector which evaluates the h_{jk} at $\boldsymbol{x}_{jk}$. The model in (4.3) is called the semi-parametric GAMLSS model. The (4.3) model is a special case of (4.1). If $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$ and $\gamma_{jk} = \boldsymbol{h}_{jk} = h_{jk}(\boldsymbol{x}_{jk})$ for specific combinations of j and k in (4.1) model, the final model includes parametric, non-parametric and random-effect terms.

The first two population parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in (4.1) model are the location and scale parameters.

For a plethora of families of population distributions a maximum of two shape parameters $\boldsymbol{\nu}(= \boldsymbol{\theta_3})$ and $\boldsymbol{\tau}(= \boldsymbol{\theta_4})$ suffice, returning the model

$$\begin{cases} g_1(\boldsymbol{\mu}) = \eta_1 = \boldsymbol{X_1}\boldsymbol{\beta_1} + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1}\boldsymbol{\gamma}_{j1} \\ g_2(\boldsymbol{\sigma}) = \eta_2 = \boldsymbol{X_2}\boldsymbol{\beta_2} + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2}\boldsymbol{\gamma}_{j2} \\ g_3(\boldsymbol{\nu}) = \eta_3 = \boldsymbol{X_3}\boldsymbol{\beta_3} + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3}\boldsymbol{\gamma}_{j3} \\ g_4(\boldsymbol{\tau}) = \eta_4 = \boldsymbol{X_4}\boldsymbol{\beta_4} + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4}\boldsymbol{\gamma}_{j4} \end{cases} \tag{4.4}$$

The (4.1) GAMLSS model is more general than the GLM, and GAM due to the fact that the distribution of the response variable is expanded over the exponential distribution family and all of the model parameters are modelled in terms of fixed and random effects.

4.1.2 Model Estimation

Essential to the manner that additive components are fitted within the GAMLSS perspective is the backfitting algorithm and the fact that quadratic penalties in the likelihood stem from the normality assumption of the random effect in the linear predictor. The estimation makes use of smoothing matrices within a backfitting algorithm.

In model (4.1) assume that the γ_{jk} have independent normal (prior) distributions with $\gamma_{jk} \sim N_{qjk}(\boldsymbol{0}, \boldsymbol{G}_{jk}^{-})$, where $\boldsymbol{G}_{jk}^{-}$ is the (generalised) inverse of a $q_{jk} \times q_{jk}$ symmetric matrix $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$; it depends on a vector of $\boldsymbol{\lambda}_{jk}$, and if $\boldsymbol{G}_{jk}$ is singular then γ_{jk} is comprehended to have an improper prior density function proportional to $exp(-\frac{1}{2}\gamma_{jk}^T \boldsymbol{G}_{jk}\gamma_{jk})$. For annotation simplicity, it is referred to $\boldsymbol{G}_{jk}$ instead of $\boldsymbol{G}_{jk}(\lambda_{jk})$ from now on.

The posterior mode estimation (or maximum a posteriori; (MAP) estimation) for the parameter vectors $\boldsymbol{\beta}_k$ and the random effect components γ_{jk} for $j = 1, \ldots, J_k$ and $k = 1, \ldots, p$ is shown to equal to the penalised likelihood estimation (Rigby and Stasinopoulos, 2005). For fixed $\boldsymbol{\lambda}_{jk}$ the $\boldsymbol{\beta}_k$ and the γ_{jk} are estimated under the GAMLSS framework by maximising a penalised likelihood estimation function l_p with the formula

$$l_p = l - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J}{}_k\gamma_{jk}^T \boldsymbol{G}_{jk}\gamma_{jk}, \tag{4.5}$$

where $l = \sum_{i=1}^{n} log\{(f(y_i|\boldsymbol{\theta}^i)\}$ is the log-likelihood function of the given data $\boldsymbol{\theta}^i$ for $i = 1, \ldots, n$. This is equivalent to the maximisation of the extended or hierarchical likelihood defined by

$$l_h = l_p + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J}{}_k\{log|\boldsymbol{G}_{jk} - q_{jk}log(2\pi)\}$$

((Pawitan, 2001, p. 429), and Lee and Nelder (1996)).

The maximisation of l_p is accomplished by the deployment of CG algorithm (Rigby and Stasinopoulos, 2005, Appendix C), which in turn leads to the shrinking (smoothing) matrix S_{jk} applied to partial residuals ϵ_{jk} to update the estimate of the additive predictor $\boldsymbol{Z}_{jk}\gamma_{jk}$ within a backfitting algorithm, given by

$$\boldsymbol{S}_{jk} = \boldsymbol{Z}_{jk}(\boldsymbol{Z}_{jk}^T \boldsymbol{W}_{kk}\boldsymbol{Z}_{jk} + \boldsymbol{G}_{jk})^{-1}\boldsymbol{Z}_{jk}\boldsymbol{W}_{kk}, \tag{4.6}$$

for $j = 1, \ldots, J_k$ and $k = 1, \ldots, p$, where $\boldsymbol{W}_{kk}$ is a diagonal matrix of iterative weights. Different forms of the matrices $\boldsymbol{Z}_{jk}$ and $\boldsymbol{G}_{jk}$ correspond to different alikes of additive terms. When it comes to random effect terms, $\boldsymbol{G}_{jk}$ is a simple and/or low order matrix, while for a cubic smoothing spline $\gamma_{jk} = \boldsymbol{h}_{jk}$, $\boldsymbol{Z}_{jk} = \boldsymbol{I}_n$ and $\boldsymbol{G}_{jk} = \lambda_{jk}\boldsymbol{K}_{jk}$, where $\boldsymbol{K}_{kk}$ is a structured matrix. In any case, the update of the term $\boldsymbol{Z}_{jk}\gamma_{jk}$ is made easy.

The $\boldsymbol{\lambda}$ hyperparameters can be either fixed or estimated; four alternative manners are suggested to avoid integrating out the random effect terms in section 4.5.4.

4.2 The Linear Predictor

4.2.1 Parametric terms

In the model (4.1) the linear predictors η_k, $k = 1, \ldots, p$ account for the parametric term $\boldsymbol{X}_k\boldsymbol{\beta}_k$ and for the additive components $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$, $j = 1, \ldots, J_k$. The parametric component contains linear and interaction effect terms for the independent variables of the model and factors, polynomials, fractional polynomials (Royston and Altman, 1994) and piecewise polynomials for variables (Smith (1979), Stasinopoulos and Rigby (1992)).

The non-linear terms included in the GAMLSS (4.1) model are estimated by either one of the following two methods:

1. the profile likelihood method

2. the derivative method

In the first method, the estimation of the non-linear parameters is accomplished by maximising the non-linear terms profile likelihood function. In the derivative method, the derivatives of the predictor $\boldsymbol{\eta}_k$ in regards of the non-linear terms are contained in the design matrix $\boldsymbol{X}_k$.

4.2.2 Additive Terms

The additive components $\boldsymbol{Z}_{jk}\boldsymbol{\gamma}_{jk}$ in (4.1) model explain a variety of terms; random effect terms smoothing and time series terms, as well. For annotation simplicity, the indices j and k in vectors and matrices where needed.

Cubic Smoothing Spline Components

It is assumed that the functions $h(t)$ are arbitrary twice continuously differentiable functions with respect to the cubic smoothing splines terms, hence the study goes on with the maximisation of the penalised log-likelihood, given by l subject to penalty terms of $\lambda \int_{-\infty}^{\infty} h''(t)^2 dt$. From Reinsch (1967), the functions $h(t)$ are all natural cubic splines, therefore they can all be expressed as linear combinations of their natural cubic spline basis functions $B_i(t)$ for $i = 1, \ldots, n$. Let $\boldsymbol{h} = h(\boldsymbol{x})$ be a vector which contains the values of the function $h(t)$ evaluated at $\boldsymbol{x}$. Let $\boldsymbol{N}$ be an $n \times n$ non-singular matrix with columns equal to the n-vectors of evaluations of functions $B_i(t)$, for $i = 1, \ldots, n$ at $\boldsymbol{x}$. Hence, $\boldsymbol{h}$ can be written with the help of the coefficient vector $\boldsymbol{\delta}$ as a linear combination of the columns of $\boldsymbol{N}$ by $\boldsymbol{h} = \boldsymbol{N}\boldsymbol{\delta}$. Assume $\boldsymbol{\Omega}$ a $n \times n$ matrix of inner products of the second derivatives of the natural cubic spline basis functions, with (r, s) given by

$$\Omega_{rs} = \int B_r''(t) B_s''(t) dt.$$

The penalty term is given by

$$Q(\boldsymbol{h}) = \lambda \int_{-\infty}^{\infty} h''(t)^2 dt = \lambda \boldsymbol{\delta}^T \boldsymbol{\Omega} \boldsymbol{\delta} = \lambda \boldsymbol{h}^T \boldsymbol{N}^{-T} \boldsymbol{\Omega} \boldsymbol{N}^{-1} \boldsymbol{h} = \lambda \boldsymbol{h}^{\boldsymbol{T}} \boldsymbol{K} \boldsymbol{h},$$

where $\boldsymbol{K} = \boldsymbol{N}^{-\boldsymbol{T}} \boldsymbol{\Omega} \boldsymbol{N}^{-1}$ is a known penalty matrix that relies solely on the values of the explanatory vector $\boldsymbol{x}$ (Hastie and Tibshirani, 2017, Ch. 2).

The model can be written in the form of the random effects GAMLSS (4.1) setting $\boldsymbol{\gamma} = \boldsymbol{h}$, $\boldsymbol{Z} = \boldsymbol{I}_n$, $\boldsymbol{K} = \boldsymbol{N}^{-\boldsymbol{T}} \boldsymbol{\Omega} \boldsymbol{N}^{-1}$ and $\boldsymbol{G} = \lambda \boldsymbol{K}$ as to $\boldsymbol{h} \sim N_n(0, \lambda^{-1} \boldsymbol{K}^{-})$; an improper prior (Silverman, 1985). It results in assuming complete prior uncertainty for higher order functions.

Parametric time series components and smoothness priors

Let an explanatory variable X have equally spaced observations x_i, $i = 1, \ldots, n$, ordered in the sequence $x_{(1)} < \cdots < x_{(i)} < \cdots < x_{(n)}$ defining an equidistant grid. Customarily, in a parametric driven time series term, X stands for time units such as days, weeks, months or years. To begin with, first and second order random walks; rw(1) and rw(2), are defined respectively by $h[x_{(i)}] = h[x_{(i-1)}] + \epsilon_i$ and $h[x_{(i)}] = 2h[x_{(i-1)}] - h[x_{(i-2)}] + \epsilon_i$, with independent errors, $\epsilon_i \sim N(0, \lambda^{-1})$, for $i > 1$ and $i > 2$, respectively, together with diffuse uniform priors for $h[x_{(1)}]$ for rw(1) and $h[x_{(2)}]$ for rw(2). Let $\boldsymbol{h} = h(\boldsymbol{x}) \Rightarrow \boldsymbol{D}_1\boldsymbol{h} \sim N_{n-1}(0, \lambda^{-1}\boldsymbol{I})$ and $\boldsymbol{D}_2\boldsymbol{h} \sim N_{n-2}(0, \lambda^{-1}\boldsymbol{I})$, where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are first and second differences matrices, respectively. The aforementioned terms can be easily included in the GAMLSS context by setting $\boldsymbol{Z} = \boldsymbol{I}_n$ and $\boldsymbol{G} = \lambda\boldsymbol{K}$ so that $\boldsymbol{\gamma} = \boldsymbol{h} \sim N(0, \lambda^{-1}\boldsymbol{K}^{-})$, where $\boldsymbol{K}$ has the form $\boldsymbol{K} = \boldsymbol{D}_1^T\boldsymbol{D}_1$ or $\boldsymbol{K} = \boldsymbol{D}_2^T\boldsymbol{D}_2$ for rw(1) and rw(2), respectively.

Penalised Splines

In smoother functions with a number of basis functions less than the number of observations, but with their regression coefficients penalised, are called penalised splines or P-splines. Eilers and Marx (1996) applied a set of q B-spline basis functions in the explanatory X. They recommended the application of a moderately large number of equally spaced knots at which the spline segments connect to guarantee enough versatility in the fitted curves. They imposed penalties on the B-spline basis function terms γ to ensure ample smoothness of the fitted curves instead. On top of that, $\boldsymbol{D}_r\boldsymbol{\gamma} \sim N_{n-r}(0, \lambda^{-1}\boldsymbol{I})$ was assumed with $\boldsymbol{D}_r$ be a $(q-r) \times q$ matrix giving r-th differences of the q-dimensional vector $\boldsymbol{\gamma}$. From the GAMLSS perspective, this equals to $\boldsymbol{G} = \lambda\boldsymbol{K}$ so that $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \lambda^{-1}\boldsymbol{K}^{-})$ where $\boldsymbol{K} = \boldsymbol{D}_r^T\boldsymbol{D}_r$.

Varying-coefficient components

Varying-coefficient models (Hastie and Tibshirani, 1993) permit the interaction between smoothing additive terms and continuous variables or factors. The incorporation to the GAMLSS framework can be easily performed with the assistance of a smoothing matrix in the form of (4.6); backfitting algorithm, but assuming that the values of R are distinct with the diagonal matrix of iterative weights $\boldsymbol{W}$ multiplied by $diag(r_1^2, \ldots, r_n^2)$ and the residuals ϵ_i for $i = 1, \ldots, n$.

Covariate random effect terms

Besag et al. (1991) and Besag and Higdon (1999) studied models for spatial (covariate) random effects with singular multivatiate normal distributions, while Breslow and Clayton (1993), Lee and Nelder (2001b) and Fahrmeir and Lang (2001) entangled these covariate terms in the predictor of the GLMMs. In model (4.1) the covariate terms can be included in the predictor of one or more of the location, scale and shape parameters.

Specific random effect terms

A plethora of random effect terms in the predictors in model (4.1) can be incorporated into, as the following:

1. overdispersion term: let $\boldsymbol{Z} = \boldsymbol{I}_n$ and $\boldsymbol{\gamma} \sim N_n(\boldsymbol{0}, \lambda^{-1}\boldsymbol{I}_n)$; provides an overdispersion term for each observation in the predictor.

2. one-factor random effect term: in model (4.1) let $\boldsymbol{Z}$ be a design matrix with elements $z_{it} = 1$ if the i-th observation belongs to the t-th factor level, otherwise $z_{it} = 0$ with $\boldsymbol{\gamma} \sim N_q(\boldsymbol{0}, \lambda^{-1}\boldsymbol{I}_q)$; provides on one factor random effect model

3. correlated random effect term: in (4.1) model correlated structure is applicable to the random effects by an appropriate selection of the matrix $\boldsymbol{G}$ due to the fact that $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{G}^{-})$

4.2.3 Terms Combinations

A combination of both the parametric and additive terms can be proposed to introduce a more complex model.

Random effects combinations

Two-level longitudinal repeated measurement design Assume a two-level design matrix with subjects as the first level in which y_{ij}, $i = 1, \ldots, n_j$, are repeated measurements at the second level on subject j, $j = 1, \ldots, J$. Let $\boldsymbol{\eta}$ be a vector of predictor values, divided into values for each subject and $\boldsymbol{Z}_j$ be an $n \times q_j$ design matrix with non-zero values for the n_j rows corresponding to j subject and assume that the $\boldsymbol{\gamma}_j$ are independent and normally distributed with distribution $N_{qj}(0, \boldsymbol{G}_j^{-1})$, for $j = 1, \ldots, J$.

Repeated measures with correlated random effect terms Set $q_j = n_j$ and $\boldsymbol{Z}_j = \boldsymbol{I}_{nj}$ for every j. A suitable choice of the matrix $\boldsymbol{G}_j$ will enable covariance or correlation structures for the random effects to be applied.

Random coefficient terms Set $q_j = q$ and $\boldsymbol{G}_j = \boldsymbol{G}$ for $j = 1, \ldots, J$. Set the non-zero submatrix of the design matrices $\boldsymbol{Z}_j$ by using the covariate(s). In return the specification of the random coefficient models is feasible.

Multilevel (nested) hierarchical model terms May each level of the hierarchy be a one-factor random effect term as in section (4.2.2.6) point (2).

Crossed random effect terms May each of the crossed factors be a one-factor random effect term as in section (4.2.2.6) point (2).

Combinations of random effects and splines

There is a number of combinations; combining random coefficients and cubic smoothing spline terms in the same covariate.

Combinations of spline terms

The combination of cubic smoothing spline in different covariates returns the additive model as introduced in Hastie and Tibshirani (2017).

4.3 Families of population distribution

4.3.1 Introduction

The population probability (density) function $f(y|\boldsymbol{\theta})$ in model (4.1) is left general without any conditional distributional form of the response variable y. The only restriction that the R implementation of a GAMLSS contains for clarifying the distribution of y is that the $f(y|\boldsymbol{\theta})$ and its first derivatives with respect to each of the parameters have to be computable. In the following table, 4.1, a plethora of multiparametric distributions studied are presented. The distributions mentioned above are used in a variety of parameterisation forms. The notation shall be used

$$y \sim D\{g_1(\theta_1) = t_1, g_2(\theta_2) = t_2, \ldots, g_p(\theta_p) = t_p\}$$

to identify uniquely a GAMLSS, where D stands for the response distribution function, $(\theta_1, \ldots, \theta_p)$ are the parameters of D, $(g_1, \ldots, g_p)$ are the link functions and $(t_1, \ldots, t_p)$ are the model formulae for the explanatory terms and/or random effects in the predictors $(\eta_1, \ldots, \eta_p)$, respectively.

$$z = \begin{cases} \frac{1}{\sigma\nu}\{(\frac{y}{\mu})^{\nu} - 1\} & if \nu \neq 0 \\ \frac{1}{\sigma} log\Big(\frac{y}{\mu}\Big) & if \nu = 0 \end{cases} \tag{4.7}$$

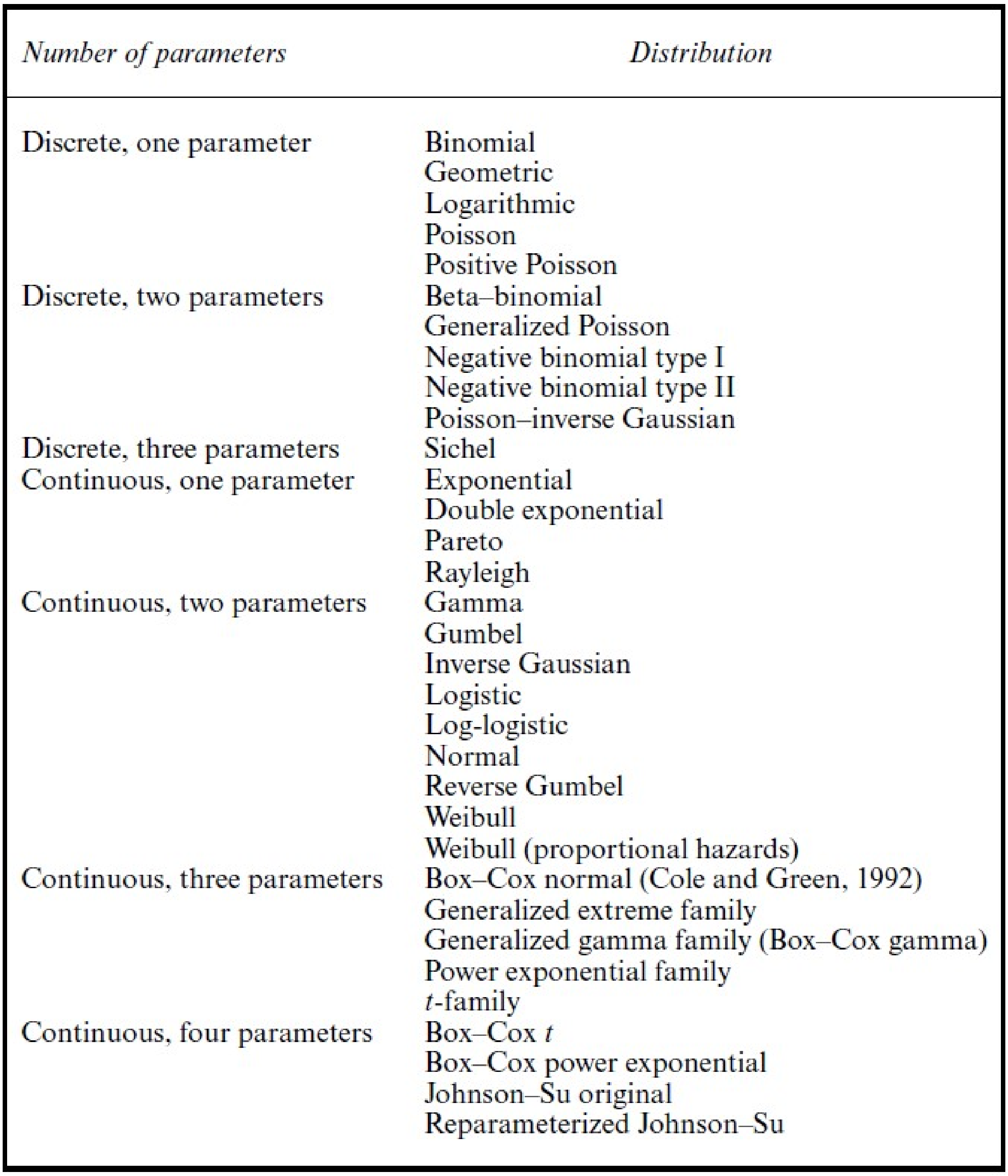
The quantile residuals can be derived once the cumulative distribution function is computed, centile estimates are accomplished given the inverse CDF can be computed. It holds for the continuous distribution functions in equation (4.7), which turn to simple standard distributions, while the CDF and inverse CDF of the discrete can be numerically approximated.

In the GAMLSS framework, censoring can be easily incorporated. To elaborate, consider that an observation is randomly right censored at value y; its contribution is given by $log\{1 - F(y|\boldsymbol{\theta})\}$, where $F(y|\boldsymbol{\theta})$ is the CDF of y. Therefore, it is sensible to demand functions for computing $F(y|\boldsymbol{\theta})$ and its first derivatives. By the same token, truncated distributions can be easily embedded into the GAMLSS.

4.3.2 Specific Distributions

There are three- and four- parameter distribution families of continuous distributions for y; defined by assuming that a transformed variable z has a well-known distribution.

The Box-Cox normal family for $y > 0$, used by Cole and Green (1992); $BCN(\mu, \sigma, \nu)$ assumes that $z \sim N(0, 1)$ where



Number of parameters	*Distribution*
Discrete, one parameter	Binomial
	Geometric
	Logarithmic
	Poisson
	Positive Poisson
Discrete, two parameters	Beta–binomial
	Generalized Poisson
	Negative binomial type I
	Negative binomial type II
	Poisson–inverse Gaussian
Discrete, three parameters	Sichel
Continuous, one parameter	Exponential
	Double exponential
	Pareto
	Rayleigh
Continuous, two parameters	Gamma
	Gumbel
	Inverse Gaussian
	Logistic
	Log-logistic
	Normal
	Reverse Gumbel
	Weibull
	Weibull (proportional hazards)
Continuous, three parameters	Box–Cox normal (Cole and Green, 1992)
	Generalized extreme family
	Generalized gamma family (Box–Cox gamma)
	Power exponential family
	t-family
Continuous, four parameters	Box–Cox *t*
	Box–Cox power exponential
	Johnson–Su original
	Reparameterized Johnson–Su

Figure 4.1: Implemented GAMLSS distributions

$$z = \begin{cases} \dfrac{1}{\sigma\nu}\left\{\left(\dfrac{y}{\mu}\right)^{\nu} - 1\right\}, & \text{if } \nu \neq 0, \\ \dfrac{1}{\sigma}\log\left(\dfrac{y}{\mu}\right), & \text{if } \nu = 0. \end{cases}$$

Cole and Green (1992) first modeled all three parameters of a distribution as non-parametric smooth functions of a single explanatory variable.

Lopatatzidis and Green (2000) reparameterised the generalised gamma family for $y > 0$; denoted as $GG(\mu, \sigma, \nu)$, assumes that $z \sim GA(1, \sigma^2\nu^2)$ where $z = (y/\mu)^{\nu}, \ \nu > 0$.

The power exponential family for $-\infty < y < \infty$ used by Nelson (1991); denoted by $PE(\mu, \sigma, \nu)$, assumes that $z \sim GA(1, \nu)$ where

$$z = \frac{\nu}{2}\left|\frac{y-\mu}{\sigma\, c(\nu)}\right|^{\nu},$$

$$c(\nu) = \left\{2^{-2/\nu}\frac{\Gamma(1/\nu)}{\Gamma(3/\nu)}\right\}^{1/2},$$

from Nelson (1991).

The Student t-family for $-\infty < y < \infty$; denoted by $TF(\mu, \sigma, \nu)$, assumes that z has a standard normal t-distribution with ν degrees of freedom; $z = \frac{y-\mu}{\sigma}$.

The four parameter Box-Cox t-family for $y > 0$; denoted by $BCT(\mu, \sigma, \nu, \tau)$, is defined by making the assumption that z has a standard t-distribution with τ degrees of freedom, Rigby and Stasinopoulos (2004a).

The Box-Cox power exponential family for $y > 0$; denoted by $BCPE(\mu, \sigma, \nu, \tau)$, is defined by making the assumption that z has a standard power exponential distribution, Rigby and Stasinopoulos (2004b). This particular distribution can be pretty handy in modelling skewness with kurtosis in the continuous data.

The Johnson-Su family for $-\infty < y < \infty$; denoted by $JSU_0(\mu, \sigma, \nu, \tau)$ (Johnson, 1949), is defined by making the assumption that $z = \nu + \tau sinh^{-1}\{(y - \mu)/\sigma\} \sim N(0, 1)$.

4.4 The Algorithms

There are two fundamental algorithms for the maximisation of the penalised likelihood in (4.5); RS and CG (Rigby and Stasinopoulos, 2005, Appendix B). The CG algorithm; generalisation of the Cole and Green (1992) algorithm, uses the first, second and cross derivatives of the likelihood function $f(y|\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ parameters being information orthogonal; the expected values of the cross-derivatives of the likelihood function are 0, for instance location and scale models and dispersion family models. In that particular case, the RS algorithm; Rigby and Stasinopoulos (1996b), for fitting mean and dispersion additive models is more appropriate. The parameters $\boldsymbol{\theta}$ are fully information orthogonal explicitly for the negative-binomial, gamma, inverse Gaussian, logistic and, normal distribution in (4.7). On that note, the RS algorithm used has met great success in fitting all the distributions in (4.7), even though it slowly converges occasionally.

The CG algorithm's main objective is to maximise the penalised likelihood function l_p for fixed hyperparameters $\boldsymbol{\lambda}$.

The major pros of these two algorithms are:

1. the modular fitting procedure; different model diagnostics for different distributions

2. easy addition of extra distributions

3. easy addition of extra additive terms

4. easily identified initial values on the grounds that starting values for the parameters $\boldsymbol{\theta}$ are only required instead of the $\boldsymbol{\beta}$ parameters.

The algorithms have been shown to be stable and fast converging for simple initial values for the $\boldsymbol{\theta}$ parameters.

In essence, for a specific data set and model the penalised likelihood can have multiple maxima. This can be easily examined by using different initial values and has generally not been found to pose an issue in the empirical application section mostly due to the large sample size.

4.5 Model Selection

4.5.1 Model Building

Let $M = \{D, G, T, \boldsymbol{\lambda}\}$ resemble the GAMLSS, where

1. D specifies the response distribution

2. G specifies the set of link functions $(g_1, \ldots, g_n)$ for parameters $(\theta_1, \ldots, \theta_n)$

3. T specifies the set of predictor terms $(t_1, \ldots, t_n)$ for predictors $(\eta_1, \ldots, \eta_p)$

4. $\boldsymbol{\lambda}$ specifies the hyperparameters set.

The GAMLSS model building process consists of comparing numerous competing models for which a variety of components $M = \{D, G, T, \boldsymbol{\lambda}\}$ combinations is tested.

Statistical inference about quantities of interest can be made either conditionally on a final model or by averaging between selected models. Inference based on a single model by conditioning is highly criticised by Draper (1995) and Madigan and Raftery (1994) based on the argument that it ignores model uncertainty and leads to underestimation of variables of interest. Underestimation reduction can be reduced by averaging between chosen models (Hjort and Claeskens, 2003).

4.5.2 Model selection; inference and diagnostics

For parametric GAMLSS, each model M of the form (4.2) is assessed by its global deviance; $GD = -2l(\hat{\boldsymbol{\theta}})$ where $l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} l(\hat{\boldsymbol{\theta}}^{i})$. Consider two nested parametric models M_1 and M_2, with global deviances GD_0 and GD_1 and error degrees of freedom df_{e0} and df_{e1}, respectively, can be compared using the generalised likelihood ratio test statistic $\Lambda = GD_0 - GD_1$ which follows an asymptotic χ^2 distribution with $d = df_{e0} - df_{e1}$ degrees of freedom, under M_0. For every model M the error degrees of freedom is defined as $df_e = n - \sum_{k=1}^{p} df_{\theta_k}$.

As far as model comparison of non-nested GAMLSS models is concerned, the generalised Akaike information criterion, GAIC, can be of use to impose a penalty on overfitted models. It is obtained by adding to the global deviance a penalty for each effective degree of freedom used in the model. The model that scores the lowest GAIC is selected among the proposed ones. The two widely known criteria; AIC and the Schwartz Bayesian information criteria SBC are special cases of the GAIC. These two criteria are asymptotically justified predicting the degree of fit in new data sets. Applying GAIC allows for the specification of different penalties that can be imposed based on modelling objectives.

The hyperparameters $\boldsymbol{\lambda}$ of GAMLSS models can be estimated with a variety of techniques; see section 4.5.4. Random effect models can be compared using their maximised profile marginal likelihood of $\boldsymbol{\lambda}$ (eliminating fixed and random effects). On top of that, different fixed effects models can be compared by using their approximate maximised marginal likelihood of $\boldsymbol{\beta}$ evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and l_h conditional on chosen parameters.

In the case of testing for a specific fixed effect predictor parameter is different from 0, a parametric χ^2 test is used comparing the change in global deviance Λ for parametric models; the change in approximate marginal deviance (removing the random effects) for random effect models, once the parameter is set to 0 with a χ^2_1 critical value. The profile likelihood for fixed effect model parameters is taken on the construction of confidence intervals. Both the aforementioned test and the confidence intervals are conditional on any hyperparameters being fixed at specific values.

An alternative manner to address this issue is to split the data set into

1. training
2. validation
3. test set

and involve them in the model fitting, selection and assessment procedure respectively.

For each M, the residuals (normalised, randomised, quantile) are used to check the adequacy of M; the distribution component D. The residuals are obtained from $\hat{r}_i = \Phi^{-1}(u_i)$, where Φ^{-1} is the inverse CDF of a standard normal variate and $u_i = F(y_i|\hat{\boldsymbol{\theta}}_i)$ if y_i is an observation from a continuous response, while u_i is a random value from the uniform distribution on the interval $[F(y_i - 1|\hat{\boldsymbol{\theta}}^{i}), F(y_i|\hat{\boldsymbol{\theta}}^{i})]$ for y_i be an observation from a discrete integer response; $F(y|\boldsymbol{\theta})$ stands for the CDF. For a right-censored continuous response u_i is defines as random value uniformly distributed from the interval $[F(y_i|\hat{\boldsymbol{\theta}}^i), 1]$. On that note, when randomisation is used, several randomised sets of residuals (or a median set of them) should better be analysed before a final decision with respect to the model M adequacy is made. If the model is correct; quite close to the correct model, then the residuals are normally distributed.

4.5.3 Posterior Mode Estimation for $\boldsymbol{\beta}$ and random effects γ

For model (4.1) an empirical Bayesian argument is applied to derive MAP and posterior mode estimation of $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{jk}$ given normal or improper priors are assumed for random effects. It is shown that this equals to the penalised likelihood maximisation l_p. The components of a GAMLSS (4.1) are

1. $\boldsymbol{y}$ the response vector of length n
2. $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_p)$ design matrices
3. $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)$ linear predictors
4. $\boldsymbol{Z} = (\boldsymbol{Z}_{11}, \dots, \boldsymbol{Z}_{J_1 1}, \dots, \boldsymbol{Z}_{1p}, \dots, \boldsymbol{Z}_{J_p p})$ design matrices
5. $\gamma^T = (\boldsymbol{\gamma}_{11}, \dots, \boldsymbol{\gamma}_{J_1 1}, \dots, \boldsymbol{\gamma}_{1p}, \dots, \boldsymbol{\gamma}_{J_p p})$ random effects
6. $\lambda^T = (\boldsymbol{\lambda}_{11}, \dots, \boldsymbol{\lambda}_{J_1 1}, \dots, \boldsymbol{\lambda}_{1p}, \dots, \boldsymbol{\lambda}_{J_p p})$ hyperparameters

Let the joint distribution of all the components in model (4.1) be

$$f(\boldsymbol{y}, \beta, \gamma, \boldsymbol{\lambda}) = f(\boldsymbol{y}|\beta, \gamma) f(\gamma|\boldsymbol{\lambda}) f(\boldsymbol{\lambda}) f(\beta), \tag{4.8}$$

where $f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma)$ and $f(\gamma|\boldsymbol{\lambda})$ are the conditional distributions for $\boldsymbol{y}$ and $\boldsymbol{\gamma}$ and $f(\boldsymbol{\beta})$ and $f(\boldsymbol{\lambda})$ are appropriate priors for $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$, respectively. $\boldsymbol{X}$ and $\boldsymbol{Z}$ are fixed and known. Assume that the $\boldsymbol{\lambda}$ are fixed and consider a constant improper prior for $\boldsymbol{\beta}$, then the posterior distribution for $\boldsymbol{\beta}$ is

$$f(\boldsymbol{\beta}, \gamma|\boldsymbol{y}, \boldsymbol{\lambda}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma) f(\gamma|\boldsymbol{\lambda}). \tag{4.9}$$

Model (4.1) assumes independent y_i for every i and let $\boldsymbol{\gamma}_{jk}$ have independent normal, possibly improper prior distribution; $\boldsymbol{\gamma}_{jk} \sim N(0, \boldsymbol{G}_{jk}^-)$ therefore,

$$log\{f(\boldsymbol{\beta}, \gamma|\boldsymbol{y}, \boldsymbol{\lambda})\} = l_p + c(\boldsymbol{y}, \boldsymbol{\lambda}),$$

where l_p is given in (4.5) and $c(\boldsymbol{y}, \boldsymbol{\lambda})$ is a function of $\boldsymbol{y}$ and $\boldsymbol{\lambda}$. Take into account that for a GAMLSS, l_p is equivalent to the h-likelihood of Lee and Nelder (1996, 2001a,b).

Henceforth, l_p is maximised over $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ yielding posterior mode (or MAP) estimation and for fixed hyperparameters $\boldsymbol{\lambda}$, MAP estimation of $\boldsymbol{\beta}, \boldsymbol{\gamma}$ is equivalent to the penalised likelihood l_p maximisation.

4.5.4 Hyperparameter Estimation

The hyperparameter's, $\boldsymbol{\lambda}$, estimation can be performed within a classical likelihood framework for random effects by maximising the marginal likelihood for $(\boldsymbol{\beta}, \boldsymbol{\lambda})$

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\gamma}) = \int f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma} f(\boldsymbol{\gamma}|\boldsymbol{\lambda}) d\boldsymbol{\gamma}.$$

The $L(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\gamma})$ maximisation contains high dimensional integration so any attempt will be computationally intensive. The maximum likelihood estimator of $\boldsymbol{\beta}$, in general, will not necessarily be the same as the respective MAP estimator.

In restricted maximum likelihood estimation (REML) a constant prior is assumed for $\boldsymbol{\beta}$ and both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are integrated out of $f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\lambda})$ to derive the marginal likelihood $L(\boldsymbol{\lambda}|\boldsymbol{y})$; maximised over $\boldsymbol{\lambda}$.

From a fully Bayesian perspective, the inference for GAMLSS would be solely derived by applying a Monte Carlo Markov Chain iterative procedure.

The following four less computationally intensive methods are considered for hyperparamters estimation of the GAMLSS model. A summation of the methods is enclosed in the algorithm

1. Procedure 1: estimate the hyperparameters $\boldsymbol{\lambda}$ by one of the techniques
   (a) profile GAIC minimisation over $\boldsymbol{\lambda}$
   (b) profile generalised cross-validation criterion maximisation over $\boldsymbol{\lambda}$
   (c) approximate marginal density maximisation for $\boldsymbol{\lambda}$ using Laplace maximisation
   (d) approximate marginal density maximisation for $\boldsymbol{\lambda}$ using an EM algorithm
2. Procedure 2: deploy RS or CG algorithm for GAMLSS model to derive the posterior mode or MAP estimates of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$

The methods are thoroughly examined below.

Profile GAIC minimisation over $\boldsymbol{\lambda}$

Hastie and Tibshirani (2017) considered GAIC for hyperparameter estimation in GAMs. A cubic smoothing spline $h(x)$ is used to model the dependence of a predictor on explanatory variable x. For a single smoothing spline term, selection of λ may be accomplished by minimising GAIC.

For a model that contains p cubic smoothing splines in different explanatory variables, the corresponding p smoothing hyperparameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ can be jointly estimated by the minimisation of GAIC over $\boldsymbol{\lambda}$.

The GAIC criterion can be applied to estimate the hyperparameters $\boldsymbol{\lambda}$ in the distribution of random effects. The degrees of freedom for random effect models are necessary. For a model with a single random effects term can be calculated as the trace of random effect smoothers $\boldsymbol{S}$; given by equation (4.6). For smoothing terms when there are other terms in the model $\sum_{i=1}^{p} tr(\boldsymbol{S})$ is an approximation of the full model complexity degrees of freedom.

Profile generalised cross-validation criterion maximisation over $\boldsymbol{\lambda}$

The generalised cross-validation criterion was taken into account for hyperparameters estimation in GAMs. The criterion GAIC is replaced by the generalised cross-validation criterion minimised over $\boldsymbol{\lambda}$; the approximate equivalence of generalised cross-validation and REML methods of estimating $\boldsymbol{\lambda}$ in smoothing spline models.

Approximate marginal density maximisation for $\boldsymbol{\lambda}$ using Laplace maximisation

The dispersion component was estimated by Lee and Nelder (1996) using a first-order approximation to the Cox and Reid (1987) profile likelihood which in turn eliminates the nuisance parameters from the marginal likelihood; adjusted profile h-likelihood.

From a Bayesian perspective; uniform improper priors for both $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ the posterior marginal of $\boldsymbol{\lambda}$ is given by

$$f(\boldsymbol{\lambda}|\boldsymbol{y}) = \int\int \frac{exp(l_h)}{f(\boldsymbol{y})} d\boldsymbol{\gamma} d\boldsymbol{\beta}, \tag{4.10}$$

where

$$l_h = l_h(\boldsymbol{\beta}, \boldsymbol{\gamma}) = log\{f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})\} + log\{f(\boldsymbol{\gamma}|\boldsymbol{\lambda})\} = l_p + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_p}\{log|\boldsymbol{G}_{jk}| - q_{jk}log(2\pi).$$

By applying a first order Laplace approximation to the integral (4.10) it is

$$f(\boldsymbol{\lambda}|\boldsymbol{y}) \approx \frac{exp(\hat{l}_h}{f(\boldsymbol{y})}\left|\frac{\hat{\boldsymbol{D}}}{2\pi}\right|^{-1/2}, \tag{4.11}$$

where $\hat{l}_h = l_h(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$

$$\hat{\boldsymbol{D}} = \boldsymbol{D}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = -\begin{pmatrix} \frac{\partial^2 l_h}{\partial\beta\partial\beta^T} & \frac{\partial^2 l_h}{\partial\beta\partial\gamma^T} \\ \frac{\partial^2 l_h}{\partial\gamma\partial\beta^T} & \frac{\partial^2 l_h}{\partial\gamma\partial\gamma^T} \end{pmatrix}, \tag{4.12}$$

evaluated at $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ and $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}(\boldsymbol{\lambda})$. The $\boldsymbol{\lambda}$ estimation can be accomplished by maximising (4.11) over $\boldsymbol{\lambda}$. Another approach would be to consider a generalisation of REML estimation of $\boldsymbol{\lambda}$, maximising an approximate profile log-likelihood for $\boldsymbol{\lambda}$; $L(\boldsymbol{\lambda})$ obtained from replacing $\boldsymbol{D}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ by the expected information $\hat{\boldsymbol{H}} = \boldsymbol{H}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$,

$$l(\boldsymbol{\lambda}) = \hat{l}_h - \frac{1}{2}log|\hat{\boldsymbol{H}}/2\pi|. \tag{4.13}$$

Approximate marginal density maximisation for $\boldsymbol{\lambda}$ using an EM algorithm

The approximate EM algorithm is used (Fahrmeir et al. (2001, p 298-303) and Diggle et al. (2002, p 172-175)), to maximise the marginal likelihood, $L(\boldsymbol{\lambda})$, over $\boldsymbol{\lambda}$ (or the posterior marginal distribution of $\boldsymbol{\lambda}$ for a non-informative uniform prior).

During the E-step of the EM algorithm $M(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}}) = E[log\{f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\lambda}\}]$ is estimated; the expectation is over the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given $\boldsymbol{y}$ and $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$, it is

$$M(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}}) = -\frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}(tr[\boldsymbol{G}_{jk}\{\hat{\gamma}_{jk}\hat{\gamma}_{jk}^T + \hat{V}(\hat{\gamma}_{jk})\}] - log|\boldsymbol{G}_{jk}|), \qquad (4.14)$$

where $\widehat{\boldsymbol{\gamma}_{jk}}$ and $\hat{V}(\widehat{\boldsymbol{\gamma}_{jk}})$ are the posterior mode and curvature of $\boldsymbol{\gamma}_{jk}$ from the MAP estimation (Rigby and Stasinopoulos, 2005, Appendix C) .

As for the M-step, $M(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}})$ is maximised over $\boldsymbol{\lambda}$ by a numerical maximisation iterative procedure. If $\boldsymbol{G}_{jk} = \boldsymbol{G}_k$, for $j = 1, \ldots, J_k$ and $k = 1, \ldots, p$ and the $\boldsymbol{G}_k$ are unconstrained positive definite symmetric matrices then the equation (4.14) is maximised returning for $k = 1, \ldots, p$,

$$\hat{\boldsymbol{G}_k}^{-1} = \frac{1}{J_k}\sum_{j=1}^{J}{}_k\{\hat{\gamma}_{jk}\hat{\gamma}_{jk}^T + \hat{V}(\hat{\gamma}_{jk})\}. \qquad (4.15)$$

Chapter 5

Empirical Analysis of GAMLSS

In this chapter empirical analysis of data coming from the Fourth Dutch Growth Study (Fredriks et al., 2000a,b), which is a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years; height and BMI are measured among other variables for Dutch boys. The empirical study develops percentile regression curves in the context of GAMLSS models aligned with the guidance of the World Health Organization (Organization et al., 2006) for the percentile values; (3,15,50,85,97). A statistical model is developed to predict and explain the relationship between the Dutch boys age and their BMI development.

5.1 Centile Estimation

Centile estimation is predominantly used in analysing age-related data of human growth. The ordinary estimation of the centile curves pertains both the response variable and the explanatory ones. The $100p$ centile of a continuous random variable Y is the value y_p, so that $Pr(Y \leq y_p) = p$ hence $y_p = F^{-1}(p)$ and y_p is the inverse cdf of Y over p. By varying values of x, a $100p$ centile curve of $y_p(x)$ against x is obtained. Centile curves can be calculated for any value of p; in this study the WHO guidance is used (3,15,50,85,97).

On top of that, $z-$score is given by the values of y and x is set as $z = \Phi^{-1}[F_{Y|x}(y)]$, where Φ^{-1} stands for the inverse cdf of the standard normal distribution. For the values of y and x used in the estimation of a statistical model, the z-scores are the residuals of a GAMLSS fitted model.

For the creation of centile curves for the response Y against x are based on non-parametric methods on the grounds that parametric methods; polynomials or even fractional polynomials (Royston and Altman, 1994) do not allow the needed flexibility to capture the characteristics of the data. The smoothing degree in smoothing methods relies on the smoothing parameters and differs for the data set studied each time. Several techniques have been recommended which are classified as follows:

- Subjective methods: the use of prior knowledge and experience coupled with general guidelines to determine the smoothing degree and create the centile curves.
- Automatic methods: in such a procedure a criterion is applied to select the smoothing parameters; AIC or any generalisation of it.
- Diagnostics methods: diagnostics as the Buuren and Fredriks (2001) worm plot or Q-statistics of Royston and Wright (2000) can be used to determine the amount of smoothing. Poor worm plots or Q-statistics may indicate the need to decrease the values of the smoothing parameters (Rigby and Stasinopoulos, 2006).

In essence, a combination of the aforementioned procedures is a good practice. The most popular methodology for centile references for individuals from a population constitutes in two different models:

- the non-parametric quantile regression (Koenker, 2005; Koenker and Bassett Jr, 1978; Koenker and Ng, 2005; Ng and Maechler, 2007).
- the parametric LMS method developed by Cole (1988); Cole and Green (1992), and its extensions, Wright and Royston (1997); Rigby and Stasinopoulos (2004b, 2006).

The LMS method was performed so it will be briefly presented in the following section.

5.2 The LMS method

The LMS method by Cole (1988); Cole and Green (1992) and its extensions was created with the aim of building centile curves for the response variable Y against a single explanatory variable x. The method assumes that the response has a specific distribution, centile curves for every p can be obtained simultaneously. The estimation of $y_p(x)$ can be easily done using the LMS method.

The LMS approach can be integrated into the GAMLSS perspective by assuming a Box-Cox Cole and Green (BCCG) distribution for the response variable; appropriate for positively or negatively skewed data with $Y > 0$. Let the positive random variable be defined through the transformed random variable Z given by

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right], & if\ \nu \neq 0, \\ \frac{1}{\sigma}log\left(\frac{Y}{\mu}\right), & if\ \nu = 0, \end{cases} \tag{5.1}$$

where $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ and $Z \sim \textit{truncated}\ N(0,1)$. As a matter of fact, the condition $0 < Y < \infty$ leads to $-\frac{1}{\sigma\nu} < Z < \infty$ if $\nu > 0$ and $-\infty < Z < -\frac{1}{\sigma\nu}$ for $\nu < 0$ which requires the truncated standard normal of Z.

The extension of the LMS method was orchestrated by Rigby and Stasinopoulos (2004b, 2006) it models the skewness in the data, by introducing the Box-Cox power exponential distribution (BCPE) and the Box-Cox t distribution (BCT) for the response variable and named the output LMSP and LMST, respectively. The difference between the assumed distributions is that BCPE assumes for the transformed variable Z a truncated power exponential distribution, while for BCT assumes a truncated t-distribution. In GAMLSS framework the BCCG, BCPE and BCT assumptions are feasible. The GAMLSS model in the centile estimation of Y for x is

$$\begin{aligned} Y \sim & D(\mu, \sigma, \nu, \tau) \\ g_1(\mu) &= s_1(u) \\ g_2(\sigma) &= s_2(u) \\ g_1(\nu) &= s_3(u) \\ g_1(\tau) &= s_4(u) \\ u &= x^{\xi} \end{aligned} \tag{5.2}$$

where D stands for the BCCG, BCPE or BCT distribution, with μ, σ, ν and τ be the approximate median, approximate coefficient of variation, skewness and kurtosis of the distribution, respectively. The functions $g(\cdot)$ represent the non-parametric smoothing functions and ξ is the power exponent of x.

The power transformation of x is needed when the response has an early or late effect on growth. On that occasion, the power transformation can extend the x scale, improving the fit of the smooth curve.

The link function $g(\cdot)$ is performed in a fashion to ensure that the parameters are appropriately defined. The original formulation of LMS by Cole and Green (1992) uses the identity links for every parameter of BCCG, while the first formulation of BCCG, BCPE, and BCt distributions have an identity link function for μ by default. The default links for σ, ν, τ are log, identity and log, respectively. During R computation all the distributions BCCGo, BCPEo and BCTo take identity link function for μ to ensure it remains positive, but they remain the same for the rest of the parameters.

The non-parametric smoothing functions $s(\cdot)$ most often require the specification of a smoothing parameter λ or the equivalent effective degrees of freedom; Hastie and Tibshirani (2017); Wood (2017).

5.3 Model Selection for the LMS approach

The selection of link functions for μ, σ, ν, τ is not problematic. As a matter of fact, the log link is preferred for σ and τ; to ensure the parameter values remain positive, the identity link is assumed for ν; $-\infty < \nu < \infty$, and for μ is opted the log link under the distribution assumption of BCCGo, BCPEo and BCTo but the identity link would do in most cases; BCCG, BCPE and BCT. The link function that is finally the one to be selected is that particular link function for which the GAIC(k) is minimised, for a specific penalty value of k.

Given a chosen distribution in (13.2) and its chosen link functions, the model specification comes down to determining the effective degrees of freedom for modelling $\mu, \sigma.\nu, \tau$ and ξ in the power transformation of x. Therefore, five hyperparameters need to be specified $(df_\mu, df_\sigma, df_\nu, df_\tau, \xi)$.

The three main approaches with respect to hypereparameters estimation, that have been considered in the literature are presented below.

- ***Method* 1 :** The minimisation of GAIC(k) for a specific penalty value k over the five hyperparameters. The deployment of an automatic procedure relying on arithmetic optimisation for the minimisation of GAIC(k) was recommended by Rigby and Stasinopoulos (2006). The BCT distribution was used and different fixed values of the penalty term k; k=2(AIC) and k=log(n)(SBC). In the first approach, the model overfitted the data yielding erratic cenitle curves in contrast to the SBC perspective which underfitted the data leading to oversmooth; biased, centile curves and unsatisfactory residual diagnostics. In the end, $k = 3$ gave a decent comprise between these and returned smooth growth curves which tuned the degree of fitting.

- ***Method* 2 :** In this method the minimisation over the five hyperparameters of the validation global deviance(VDEV) is the objective (Stasinopoulos et al., 2007). The data set is split randomly into the training set; 60% of the data, and the validation set; 40% of the data. For each and every single set of hyperparameters the model (5.2) is fitted to the training data and the resulting validation global deviance $VDEV = -2\tilde{l}$is calculated, where $\tilde{l}$ is the log-likelihood of the validation set. VDEV is in turn minimised over the five hyperparameters. That particular method was assessed to moderately fit the data.

- ***Method* 3 :** This is a two-step approach. First, in case a transformation on the x-axis is necessary then for a normal distribution model with $g(\mu) = s_1(x^\xi)$ and constant σ; GAIC(k) is minimised over ξ for fixed penalty value k. Given the estimated ξ the second step involves the model (5.2) to be fitted for the distribution D and calculate the four degrees of freedom hyperparameters $(df_\mu, df_\sigma, df_\nu, df_\tau)$ from a local ML procedure;see Sections 3.4 and 9.4 of Rigby and Stasinopoulos (2014); Stasinopoulos et al. (2017). The distribution for which the criterion GAIC(k) is minimised is the chosen. It is considered as the fastest of the methods and returns a model with similar centiles to the two latter methods.

5.4 The Dutch Boys BMI data

The data analysed are from the Fourth Dutch Growth Study (Fredriks et al., 2000a,b). The BMI (y) is the variable of interest modelled by the age (x) of the boys. The objective is to obtain smooth reference centile curves for BMI against age.

In Figure 5.1 the BMI against age is plotted for the full data and the training data set in (a) and (b), respectively. Taking a closer look at the data sets, it is shown that there are intervals of ages in which the observations are more concentrated indicating potential positive skewness of BMI distribution, Figure 5.2, given age as well as a non-linear relationship between the location (and possibly the scale, skewness and kurtosis) of BMI with age (Rigby and Stasinopoulos, 2005). In a former study of Cole et al. (1998) in the Dutch girls BMI, similar problems were addressed (positive data skewness), using the LMS method showed significant kurtosis in the residuals after model fit drawing the conclusion that kurtosis was not adequately faced. A power transformation of age to $X = age^{\xi}$ contributes drastically to the model fit (Rigby and Stasinopoulos, 2004a).

Therefore, given $X = x$ the dependent variable BMI is modelled using a Box-Cox t-distribution; BCT(μ, σ, ν, τ). The arrival to that distribution selection comes as a result of the application of the LMS method which accomplishes the minimisation of GAIC(3) under the BCT assumption; Global Deviance(GD)=19869.23. The model parameters μ, σ, ν, τ are modelled as non-parametric smooth functions of x; given $X = x_i$ then $y_i \sim BCT(\mu_i, \sigma_i, \nu_i, \tau_i)$ independent for $i = 1, \ldots, n$, where

$$\begin{aligned} \mu_i &= h_1(x_i), \\ log(\sigma_i) &= h_2(x_i), \\ \nu_i &= h_3(x_i), \\ log(\tau_i) &= h_4(x_i) \end{aligned} \tag{5.3}$$

Hence, $h_k(x)$ are smooth functions of x for $k = 1, 2, 3, 4$ and $x_i = agei^{\xi}$ for $i = 1, \ldots, n$ with ξ be a non-linear parameter in the model. The log functions are used to ensure parameter positivity.

For the fitted model P-splines were used for each model parameter; μ, σ, ν, τ with 12.06, 6.07, 4.33 and 2.00 be their degrees of freedom, respectively. As for the coefficient of the age power transformation exponent it is estimated to be 0.36.

The fitted models for μ, σ, ν, τ for the selected model are shown in Figure 5.10. The fitted ν indicates moderate to high skewness in BMI for all ages ($\hat{\nu} < 1$), while the fitted τ indicates leptokurtosis especially in the older boys. In the top left and top right panel of Figure 5.6 the residuals are plotted against the fitted values and against their index, respectively. In the bottom left and right panel of Figure 5.6 a kernel density estimate and a QQ-plot are provided, respectively. The residuals seem to be randomly distributed in both cases, which is a good sign of model fit. Nonetheless, from the QQ-plots in Figure 5.3 and the bottom right of the Figure 5.6 it can be seen that a slightly longer lower tail and one possible outlier in the upper tail; in both cases the deviations are acceptable allowing for the flexibility of the normal distribution.

Another diagnostic tool to examine the fitted model is the Q-statistics. As a measure of goodness of fit it seems that all the Q-statistics are reasonable (not statistically significant p-values). In Figure 5.8 the Z statistics indicate an adequate model on the grounds that all the absolute values of $|Z|$ are less than 2; there are no squares within the circles. The next tool for model checking is the wormplot in Figure 5.9. The produced plot is equivalent to the normal Q-Q plot in which is demonstrated how far the ordered residuals are from their expected values; the closer to the horizontal line the closer the distribution of the residuals is to the standard normal. There is no significant departure from the elliptic curves (or a clear departure from the horizontal line) all of which vindicate the model explains quite well the data. The fitted curve to the points of the worm; cubic fit line, reflects any inadequacies in the model fit. Here, the line remains straight for the most of the data with slight departures in the tails suggesting a good model fit. Multiple worm plots are produced in Figure 5.10 as a manner to highlight failures of the model within different ranges of the explanatory. The explanatory variable, age, is split into 9 equispaced intervals in which the residuals are represented so that any problematic fit is emerged. It still seems that the model fits the data well with the exception of the upper middle panel; there is an outlier in the top left quartile of the plot.

After the diagnostic model checking, the study moves on with the creation of centile curves under the Box-Cox t-distribution for the fitted model, Figure 5.7. As a matter of fact the centile fan was adjusted to meet the WHO guidance (Organization et al., 2006) of using as percentiles the values of $(3, 15, 50, 85, 97)$; Figure 5.12. It appears that the centile curves resemble a particular pattern for different age intervals so it would be recommendable to split the main curve into two segments for a thorough examination. In this case the centile reference curves are split into two distinct age groups; 0-2 year old boys and 2-21 year old boys; Figure 5.13. To elaborate in the first age group the centile curves of the BMI growth increase quite steeply for the first 6 months of infancy while they plateau until they become 2 years old. In the second age group it appears that the BMI growth remains on average the same until their fifth year, after that an increase is recorded; steeper in the beginning (5-14yrs) and slower during the beginning of adulthood.

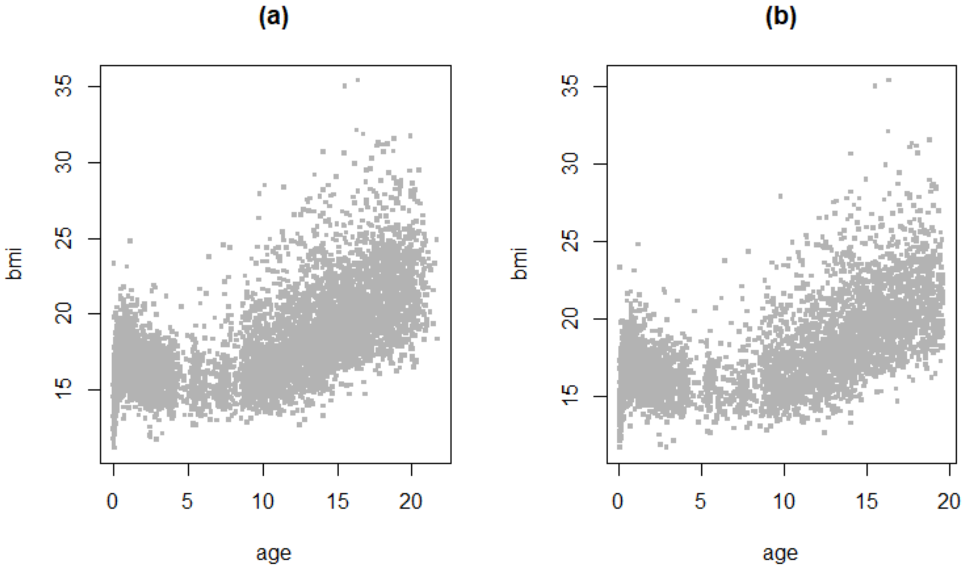


Figure 5.1: BMI against age, Dutch boys data (a) the complete data set of 7040 observations, (b) random sample of 5000 observations.

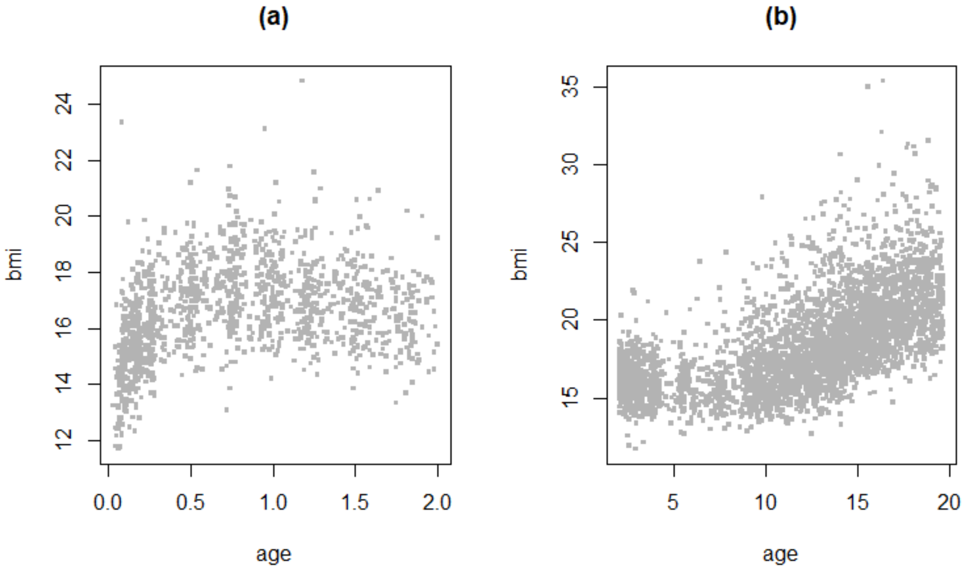


Figure 5.2: BMI against age, Dutch boys data split into 0-2 years old (a) and 2-21 years old (b)

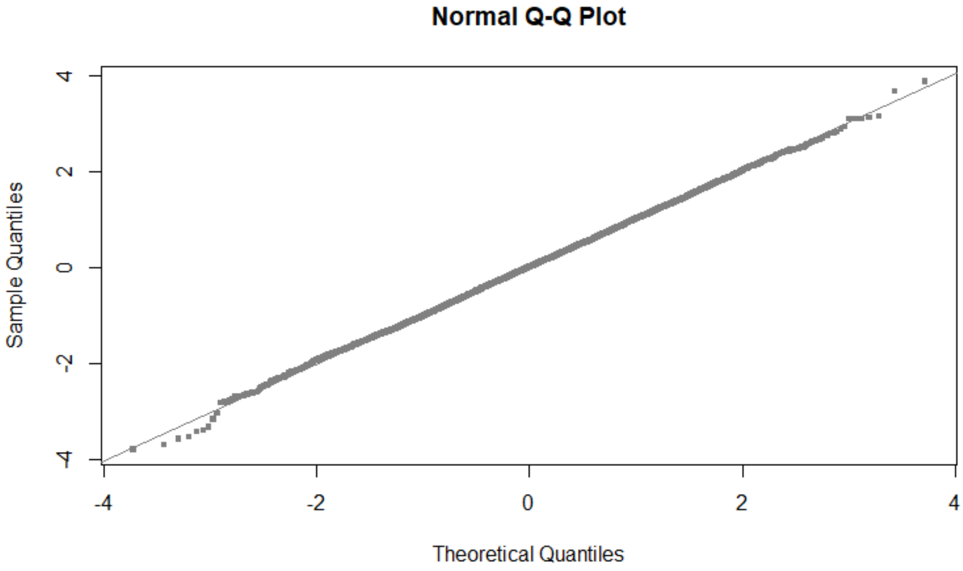


Figure 5.3: The Normal QQ-plot of sample quantiles against the theoretical quantiles

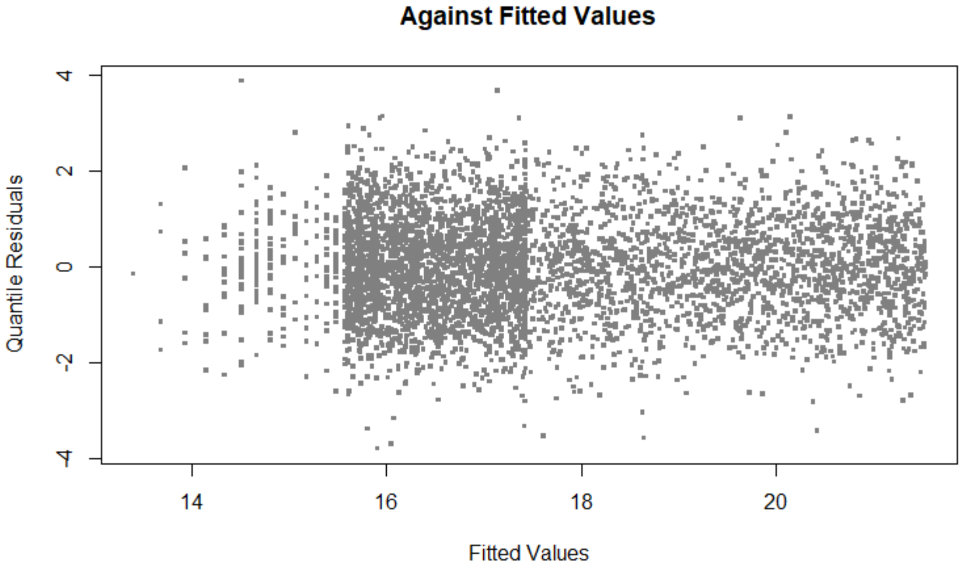


Figure 5.4: The model residuals plotted against the fitted values

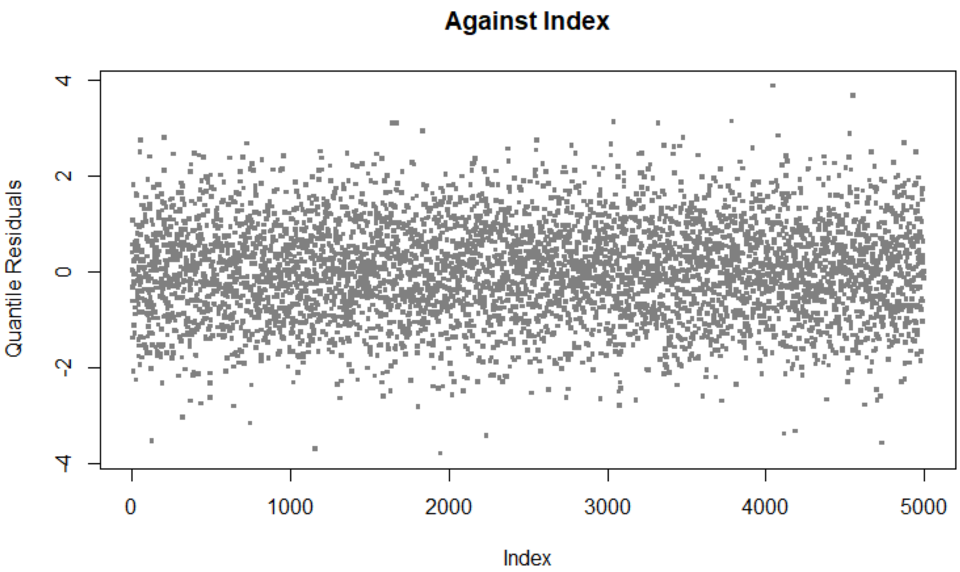


Figure 5.5: The model residuals plotted against the index

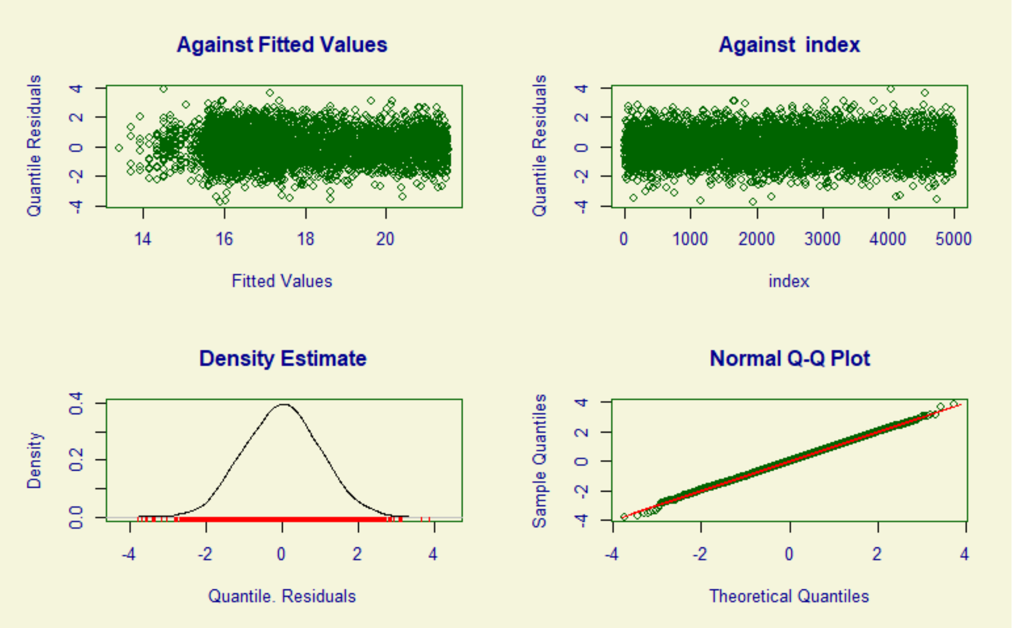


Figure 5.6: Final model Diagnostic Residuals

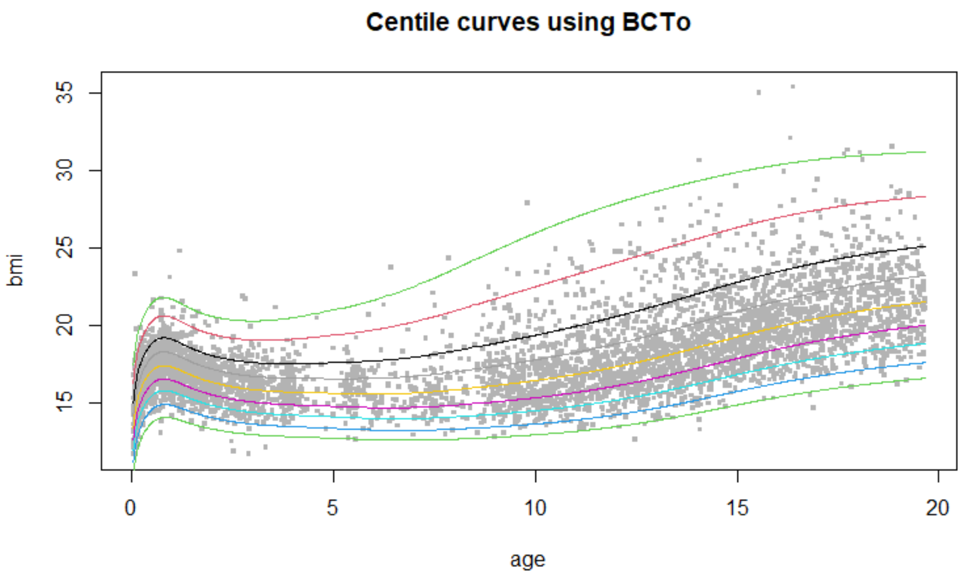


Figure 5.7: Centile Curves using the Box-Cox t-distribution (BCTo) distribution for the Dutch boys BMI

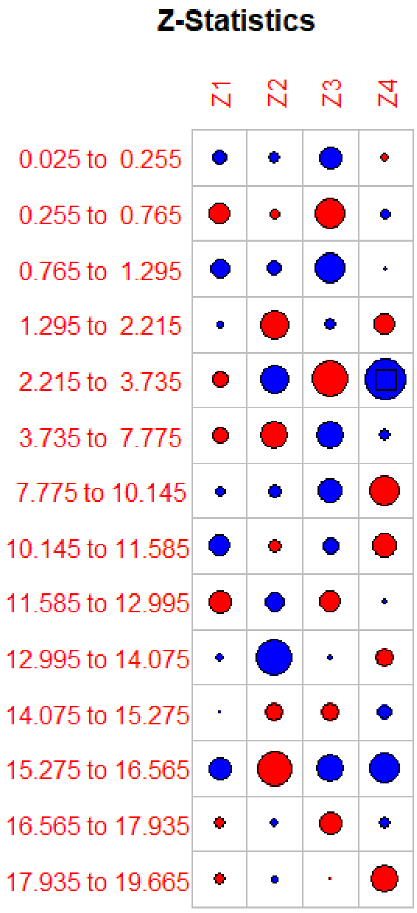


Figure 5.8: Plot of Q-statistics for the model fitted by LMS

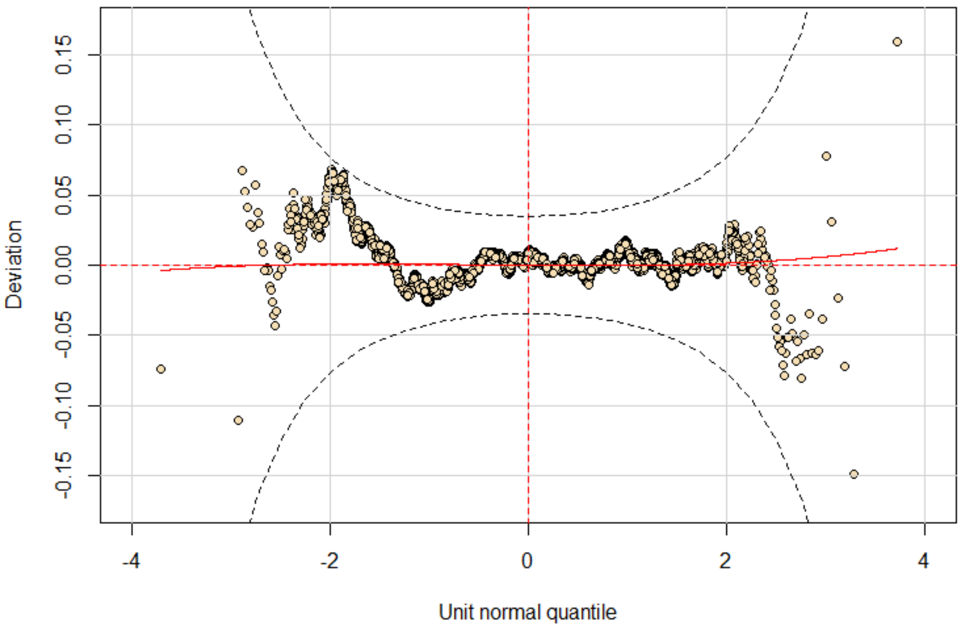


Figure 5.9: Worm plot from the BCT model

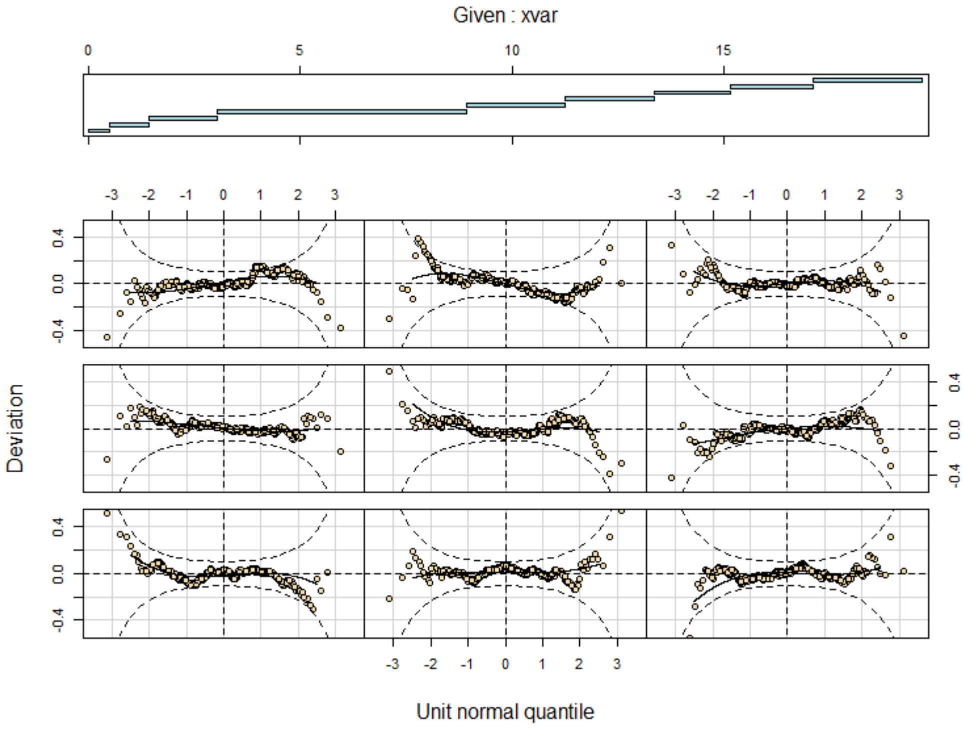


Figure 5.10: Worm plot for the fitted model by LMS

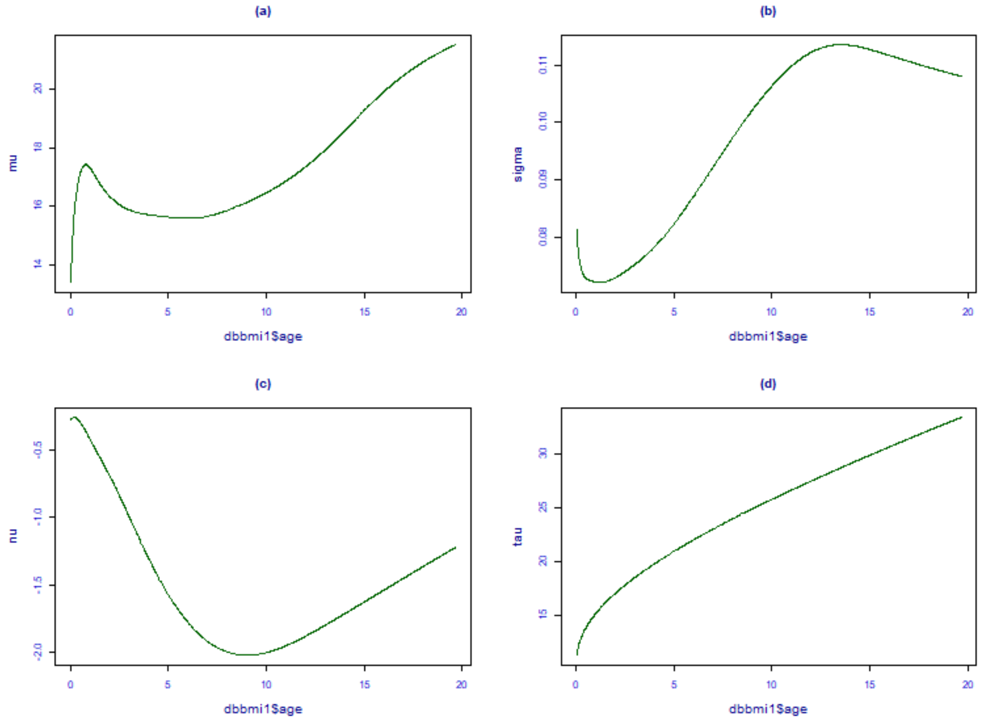


Figure 5.11: Fitted values for (a) μ (b) σ (c) ν (d) τ, against age from a Box-Cox t-distribution for Dutch boys BMI

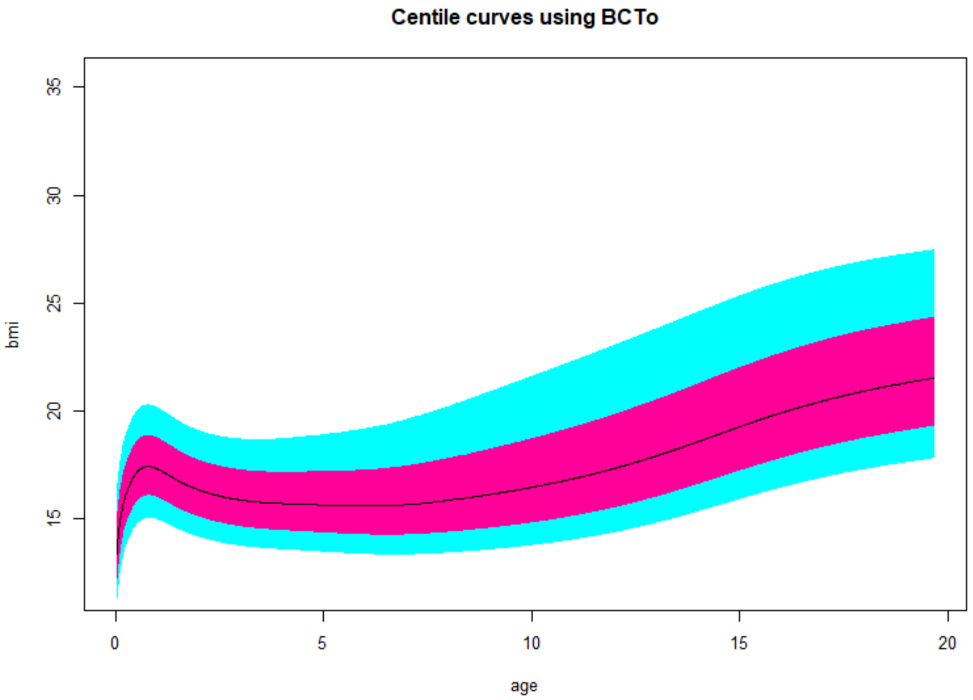


Figure 5.12: Fan-chart (centile) curves using a Box-Cox t-distribution for Dutch boys BMI

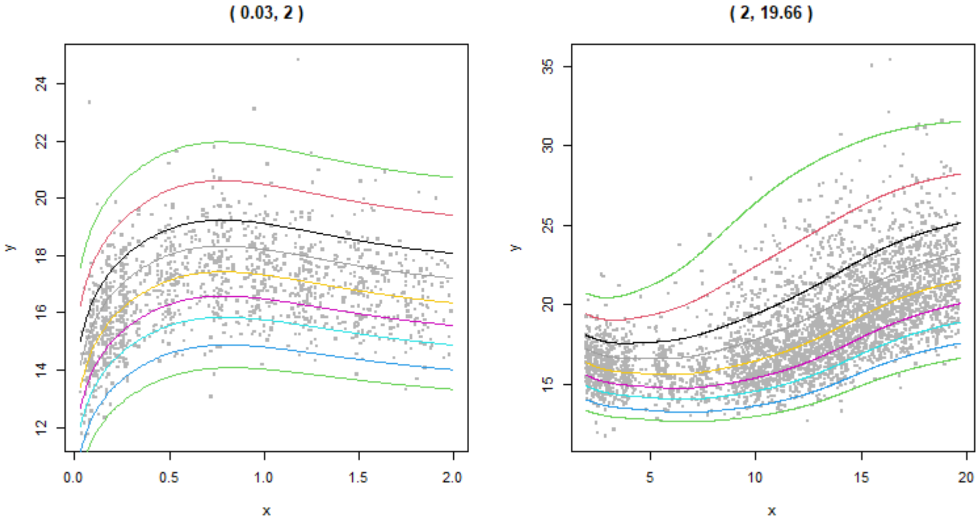


Figure 5.13: Centile curves for the two age ranges using a Box-Cox t-distribution for Dutch boys BMI

Chapter 6

Conclusion

In this dissertation numerous statistical models have been discussed. First, the generalised linear model allows for the response variable to have an error distribution that belongs to the exponential family of distributions. The generalised linear model connects the systematic part with the random component via a link function. Models for binary and count have been purposed together with estimation and inference methods.

The generalised additive model comes next, which constitutes a generalisation of the generalised linear model. In essence, the generalised additive model achieves to interconnect the properties of the generalised linear model with a set of predictor smooth functions. Multiple methods of smooth functions selection coupled with techniques to control the degree of smoothness are proposed in this particular thesis. The generalised additive models' inference criteria are presented in the end.

The final model analysed is the generalised additive model for location, scale and shape. The generalised additive model has been studied in terms of skewness and kurtosis to accomplish more flexible modelling of the data characteristics that both the previous models miss to capture. The model assumes a general distribution for all the model parameters; mean, variance, skewness and kurtosis. Its broad assumptions retain the flexibility of the model being applied to a wide range of occasions. The iterative algorithms that make possible the estimation of the model have been presented in this dissertation, as well.

Finally, an empirical application with centile estimation on real data in the framework of the generalised additive model for location, scale and shape is performed. The iterative estimation methods are applied to find the distribution function of the response variable that best fits the data, in which case is the Box-Cox t distribution, taking into account the idiosyncratic characteristics of the data with respect to mean, variance, skewness and kurtosis of the data. Centile estimates based on WHO percentile guidance are calculated, to describe the average bmi growth of the Dutch boys given their age. In essence, the centile curves are estimated under the Box-Cox t-distribution, which increase quite steeply for the first six months after birth, while for the rest months up to two years old they stabilise. The second stage of rapid bmi growth is the period of life between five to fourteen years old which slows down in the early adulthood.

Bibliography

Alan Agresti. *Foundations of linear and generalized linear models.* John Wiley & Sons, 2015.

Julian Besag and David Higdon. Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):691–746, 1999.

Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

Stef van Buuren and Miranda Fredriks. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, 20 (8):1259–1277, 2001.

Tim J Cole, Jenny V Freeman, and Michael A Preece. British 1990 growth reference centiles for weight, height, body mass index and head circumference fitted by maximum penalized likelihood. *Statistics in medicine*, 17(4):407–429, 1998.

Timothy J Cole. Fitting smoothed centile curves to reference data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 151(3): 385–406, 1988.

Timothy J Cole and Pamela J Green. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in medicine*, 11(10): 1305–1319, 1992.

David Roxbee Cox and Nancy Reid. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(1):1–18, 1987.

Robert B Davies. The distribution of a linear combination of $\chi 2$ random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):323–333, 1980.

Peter Diggle, Peter J Diggle, Patrick Heagerty, Kung-Yee Liang, Patrick J Heagerty, Scott Zeger, et al. *Analysis of longitudinal data.* Oxford university press, 2002.

David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):45–70, 1995.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121, 1996.

Ludwig Fahrmeir and Stefan Lang. Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220, 2001.

Ludwig Fahrmeir, Gerhard Tutz, Wolfgang Hennevogl, and Eliane Salem. *Multivariate statistical modelling based on generalized linear models*. Springer, 2001.

A Miranda Fredriks, Stef Van Buuren, Ruud JF Burgmeijer, Joanna F Meulmeester, Roelien J Beuker, Emily Brugman, Machteld J Roede, S Pauline Verloove-Vanhorick, and Jan-Maarten Wit. Continuing positive secular growth change in the netherlands 1955–1997. *Pediatric research*, 47(3):316–323, 2000a.

A Miranda Fredriks, Stef van Buuren, Jan M Wit, and SP Verloove-Vanhorick. Body index measurements in 1996–7 compared with 1980. *Archives of disease in childhood*, 82(2):107–112, 2000b.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Sonja Greven and Thomas Kneib. On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*, 97(4):773–789, 2010.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

Hjort and Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 2003.

Norman L Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176, 1949.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Roger Koenker and Pin Ng. Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics*, pages 418–440, 2005.

Youngjo Lee and John A Nelder. Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58 (4):619–656, 1996.

Youngjo Lee and John A Nelder. Hierarchical generalised linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88:987–1006, 2001a.

Youngjo Lee and John A Nelder. Modelling and analysing correlated non-normal data. *Statistical Modelling*, 1(1):3–16, 2001b.

Huan Liu, Yongqiang Tang, and Hao Helen Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

A Lopatatzidis and PJ Green. Nonparametric quantile regression using the gamma distribution. *submitted for publication*, 2000.

David Madigan and Adrian E Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546, 1994.

Colin L Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135 (3):370–384, 1972.

Daniel B Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, pages 347–370, 1991.

Pin Ng and Martin Maechler. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7 (4):315–328, 2007.

World Health Organization, World Health Organization, et al. Who multicentre growth reference study group: Who child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development. *Geneva: WHO*, 2007, 2006.

Yudi Pawitan. *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press, 2001.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

Robert A Rigby and D Mikis Stasinopoulos. Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution. *Statistics in medicine*, 23(19):3053–3076, 2004a.

Robert A Rigby and D Mikis Stasinopoulos. Smooth centile curves for skew and kurtotic data modelled using the box–cox power exponential distribution. *Statistics in medicine*, 23(19):3053–3076, 2004b.

Robert A Rigby and D Mikis Stasinopoulos. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):507–554, 2005.

Robert A Rigby and D Mikis Stasinopoulos. Using the box-cox t distribution in gamlss to model skewness and kurtosis. *Statistical Modelling*, 6 (3):209–229, 2006.

Robert A Rigby and Dimitrios M Stasinopoulos. Automatic smoothing parameter selection in gamlss with an application to centile estimation. *Statistical methods in medical research*, 23(4):318–332, 2014.

Robert A Rigby and DM Stasinopoulos. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing*, 6(1):57–65, 1996.

Robert A Rigby and Mikis D Stasinopoulos. Mean and dispersion additive models. In *Statistical theory and computational aspects of smoothing*, pages 215–230. Springer, 1996b.

Patrick Royston and Douglas G Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43 (3):429–453, 1994.

Patrick Royston and Eileen M Wright. Goodness-of-fit statistics for age-specific reference intervals. *Statistics in medicine*, 19(21):2943–2962, 2000.

Benjamin Saefken, Thomas Kneib, Clara-Sophie van Waveren, and Sonja Greven. A unifying approach to the estimation of the conditional akaike information in generalized linear mixed models. *Electronic Journal of Statistics*, 8(1):201–225, 2014.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

Bernhard W Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21, 1985.

Patricia L Smith. Splines as a useful and convenient statistical tool. *The American Statistician*, 33(2):57–62, 1979.

D Mikis Stasinopoulos, Robert A Rigby, et al. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, 23(7):1–46, 2007.

DM Stasinopoulos and RA Rigby. Detecting break points in generalised linear models. *Computational Statistics & Data Analysis*, 13(4):461–471, 1992.

Mikis D Stasinopoulos, Robert A Rigby, Gillian Z Heller, Vlasios Voudouris, and Fernanda De Bastiani. *Flexible regression and smoothing: using GAMLSS in R.* CRC Press, 2017.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

Grace Wahba. *Spline models for observational data.* SIAM, 1990.

Simon N Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228, 2013.

Simon N Wood. *Generalized additive models: an introduction with R.* CRC press, 2017.

Simon N Wood and Matteo Fasiolo. A generalized fellner-schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics*, 73(4):1071–1081, 2017.

Simon N Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016.

Eileen M Wright and Patrick Royston. A comparison of statistical methods for age-related reference intervals. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(1):47–69, 1997.

Dalei Yu and Kelvin KW Yau. Conditional akaike information criterion for generalized linear mixed models. *Computational Statistics & Data Analysis*, 56(3):629–644, 2012.