PERFORMANCE OF SPLINE-BASED GAM IN THE PRESENCE OF OUTLIERS AND MULTICOLLINEARITY

Master Thesis

HURUY DEBESSAY ASFHA

Eskişehir, 2017

PERFORMANCE OF SPLINE-BASED GAM IN THE PRESENCE OF OUTLIERS AND MULTICOLLINEARITY

HURUY DEBESSAY ASFHA

MASTER THESIS

Department of Statistics
Supervisor: Assoc. Prof. Dr. Betül KAN KILINÇ

Eskişehir
Anadolu University
Graduate School of Sciences
May, 2017

This thesis was supported by the project with reference number 1610F671 which was accepted by BAP commission.

ABSTRACT

PERFORMANCE OF SPLINE-BASED GAM IN THE PRESENCE OF OUTLIERS AND MULTICOLLINEARITY

HURUY DEBESSAY ASFHA

Department of Statistics
Anadolu University, Graduate School of Sciences, May, 2017
Supervisor: Assoc. Prof. Dr. Betül KAN KILINÇ

Generalized additive models (GAMs) are extension of additive models as generalized linear models (GLMs) are to ordinary linear regression model. There are different approaches of fitting these kinds of models one of which is the smoothing bases approach, where variety alternatives of smoothing functions are used to define the bases of the model matrix. Penalized regression spline which is estimated by penalized regression techniques is one alternative method for representing GAM models.

In this thesis, three penalized regression splines; cubic spline, p-spline, and thin-plate spline are proposed to fit GAM for a simulated data. The performance of these smoothers is evaluated and compared for tolerance of the effect of outliers, multicollinearity and both when they exist together. Results of the experiments showed that the GAMs fitted using these nonparametric regression techniques are less prone to multicollinearity and outliers compared to their parametric counterparts.

Keywords: Generalized additive models, Smoothing, Penalized regression spline, Outlier, Multicollinearity.

ÖZET

SPLAYN-TABANLI GAM'IN ÇOKLU BAĞLANTI VE AYKIRI DEĞER VARLIĞINDA PERFORMANSLARI

HURUY DEBESSAY ASFHA

İstatistik Anabilim Dalı
Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Mayıs, 2017
Danışman: Doç. Dr. Betül KAN KILINÇ

Doğrusal regresyon modellerinin genelleştirilmiş doğrusal modellerin bir uzantısi oması gibi genelleştirilmiş toplamsal modeller de toplamsal modellerin bir uzantısıdır. Bu tür modellerin veriye uyumu için kullanılan değişik yaklaşımlardan biri olan düzleştirme, model matrisinde çeşitli düzleştirme fonksiyonlarının kullanıldığı yaklaşımlardandır. Cezalı regresyon splaynları, genelleştirilmiş toplamsal modelleri oluşturmak cezalı regresyon tekniği ile kestirilen bir diğer yöntemdir.

Bu tez çalışmasında, üç farklı regresyon splaynları, kübik, p-splayn ve ince tabakalı splaynlar veri üretmede kullanılır. Bu düzleştiricilerin, aykırı değer, çoklu bağlantı ve her iki durum söz konusu olduğunda performansları karşılaştırılır. Sonuçlar elde edildiğinde genelleştirilmiş toplamsal modellerin, parametrik olan regresyon tekniklerine göre aykırı değer ve çoklu bağlantıdan daha az etkilendiğini ortaya konmuştur.

Anahtar Kelimeler: Genelleştirilmiş toplamsal modeller, Düzleştirme, Cezalı regresyon splaynları, Aykırı değer, Çoklu bağlantı.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my honored supervisor Assoc. Prof. Dr. Betül KAN KILINÇ for her supervision and guidance. The door to her office was always open whenever I had a question or I ran into any trouble with coding. I am extremely thankful and indebted to her for sharing her expertise, and for her sincere and valuable guidance and encouragement throughout this work.

Moreover, I would like to express my sincere gratitude to all who made contributions all the way to make these achievements happen. I am indebted to all my teachers without whom this success would have not been achieved. In particular, I would like to thank Prof. Dr. Berna YAZICI, head of the department of Statistics, for her professional advice and guidance.

I would also like to take this opportunity to express my gratitude to all my classmates for their valuable encouragement. Furthermore, my special thanks goes to my friends İsmail BAYRAM with his wife Angela ANDRADE HERNANDEZ, Clement KWANG, Ceyda ERDEM, Edward CHOME, Beldine OMONDI, Aslı KAYA, Maroua BEN GHOUL and all others for their moral and material support. Furthermore, I would like to thank the project unit of Anadolu University for funding this work.

Last but not least, heart felt thanks goes to my father Ato. Debessay Asfha and my mother Hidat Berhe and all other family members for providing me a continuous encouragement, unceasing love and unfailing support throughout my years of study. It would not have been possible to attain this achievement without them. Thank you!

Huruy Debessay Asfha

May, 2017

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND SYMBOLS

β	: Beta
δ	: Delta
η	: Eta
$\in$	: Element/s of
ϵ	: Epsilon
$\forall$	: For all
λ	: Lambda
μ	: Mu
ω	: Omega
∂	: Partial derivative
ϕ	: Phi
π	: Pi
ρ	: Rho
σ	: Sigma
θ	: Theta
C^2	: Twice differentiable
AM	: Additive Model
cr	: Cubic regression spline
GAM	: Generalized Additive Model
GCV	: Generalized Cross Validation
GLM :	: Generalized Linear Model
IRLS	: Iterative Re-weighted Least Square
LOESS	: Locally Estimated Scatterplot Smoothing
LOWESS	: Locally Weighted Scatterplot Smoothing
MGCV	: Mixed GAM Computation Vehicle
MLE	: Maximum Likelihood Estimator
MSE	: Mean Square Error
OCV	: Ordinary Cross Validation

OLS	: Ordinary Least Square
P-IRLS	: Penalized Iterative Re-Weighted Least Square
ps	: P-spline
tp	: Thin-plate spline
SVM	: Support Vector Machines

1. INTRODUCTION

Regression analysis, a technique used for investigating and modeling the relationships among variables, plays a vital role and can safely be argued that it holds a central point in statistical data analysis. Generally, regression models can be categorized as parametric, nonparametric and semiparametric (have behaviors of parametric and nonparametric). Linear regression is the basic parametric model and is written in the form of

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where, $\boldsymbol{X}$ is model matrix, $\boldsymbol{\beta}$ is a vector of unknown model coefficients, and $\boldsymbol{y}$ is response variable. $\boldsymbol{\epsilon}$ is random error term which follows $N(0, \sigma^2)$ distribution.

These kind of models are simple to use however, there are five key strict assumptions to be considered while applying linear regression models. These assumptions are:

- Linear relationship between response and predictor variables. $E(y/X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.
- Response variable is normally distributed;
  $f(y/X) \sim N(\mu, \sigma^2)$
- No or little multicollinearity among covariates
- No autocorrelation:
  $Cov(\epsilon_i, \epsilon_j) = 0$.
- Homoscedasticity;
  $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

However, these assumptions do not always hold true. Generalized linear models (GLMs) are introduced to relax the strict assumptions of normality and homoscedasticity in ordinary linear regression. In GLM, distribution of response variable has to be one of the exponential family distributions among which are normal, binomial, Poisson, exponential and gamma distributions [1, 2].

$$y \sim \textit{exponential family distributions}$$

However, the assumption of linear dependence in the classical linear models is carried over without modifications to GLMs [3]. In GLMs, the similarity in considerably several properties of the exponential family distributions allows us to use the same technique to estimate model coefficients using the likelihood concept of estimation. The general form of GLMs is [4, 3, 5],

$$g(\mu) = \eta = \boldsymbol{X}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{1.2}$$

where, g is a link function which connects the systematic component, η (called linear predictor), and the random component (response variable) of the model [6]. The expected response is then obtained as,

$$E(y) = g^{-1}(\eta) = g^{-1}(\boldsymbol{X}'\boldsymbol{\beta})$$

Here, if the response variable follows normal distribution and an identity link function is used, the generalized linear model turns out to be an ordinary linear regression model.

In parametric regression, the functional form of the model is known in advance and is then fit to the data with global estimates. In the absence of a strong evidence for the predefined functional form to represent the data appropriately, an alternative has to be used; estimating the functional form from the data is the best way to proceed [6]. In order to estimate the functional form from the data, the global estimation in parametric models has to be replaced with local estimates. These methods of locally estimating of the functional form from the data itself is generally described as nonparametric techniques [6].

Nonparametric regression models allow one to fit a flexible nonlinear model to the data in order to represent the relationship between the response and predictor variables. As parametric models, nonparametric models are useful both for modeling and diagnosis of the nonlinear relationships.

A detailed explanation of some nonparametric model fitting techniques (smoothers which apply the local estimation principle) can be found in Hastie and Tibshirani, 1990. The most popular methods of nonpara-metric regression techniques include local polynomial smoothing, kernel smoothing and splines [7, 8].

When the linear predictors in GLM model (Eq:1.2) are replaced with additive predictors, the model is called generalized additive model (GAM), thus, GAMs are regarded as extensions of GLMs.

$$g(\mu) = \eta = \beta_0 + f_1(x_1)\beta_1 + f_2(x_2)\beta_2 + \ldots + f_p(x_p)\beta_p + \epsilon \tag{1.3}$$

The f_i's in generalized additive models are smoothing functions which can be any of kernels, local regression (loess) or smoothing splines. Due to the fact that GAMs can incorporate nonparametric models into parametric ones, they can sometimes be described as semiparametric regression models [6]. In semiparametric models, some predictor variables are modeled parametrically while others are modeled using nonparametric regression. In GAMs, while the assumption of standard linear models of the linear dependency of $\boldsymbol{y}$ on $\boldsymbol{X}$ is relaxed, the additivity assumption still holds true and it is this additivity property that makes GAMs easier to interpret than other algorithms such as supportive vector machines (SVM), neural networks, ... etc. [6].

Generally, GAMs are computationally expensive techniques compared to the linear models due to the fact that they build the model using local fits. However, different algorithms have been developed to fit GAM models iteratively. The *gam* package was developed based on the work of Hastie and Tibshirani, 1990 to fit generalized additive models to the data of concern [9, 10]. This gam function constructs GAM models by combining different smoothing methods using backfitting algorithm. Another package used to fit GAM models in R is the *mgcv* package of Wood, 2006. It employs the approach of penalized regression spline to fit a model [4]. By default, the degree of smoothness of the fit is chosen internally by the algorithm. Automatic selection of smoothing parameter is an advantage for the reason that it avoids the subjectivity and work of choosing it by the user. However, it can fail to obtain the best degree of smoothness and human intervention could sometimes be needed [10].

In parametric regression analysis, there are different causes of model disturbances one of which is the existence of abnormal observations in the dataset which distorts the model parameters. Furthermore, this in turn may result in an inflated estimate of σ^2, the residual sum of squares. Similarly [11], in nonparametric regression the presence of small percentage of such anomalous observations in the data affects the estimated smooth functions causing the model to be more close to them. In literature, some outlier-resistant GAM fitting techniques have been developed. Alimadad and Salibian-Barrera, 2011 discussed a robust method of GAM fitting technique by using those derived from robust quasi-likelihood equations in place of the maximum likelihood based weights in the local scoring algorithm [11]. The *rgam* in R is an implementation of this robust method. Wong et al., 2014 proposed an M-type robust estimating technique to fit a more robust generalized additive model in the presence of outliers and was implemented in R as *robustGAM.* The core idea of this method is to decompose the overall M-type problem into a sequence of additive models fitting problems [12].

The existence of concurvity which leads to a poor estimation of model parameters and underestimation of their standard error is another cause of model disturbance. According Buja et al., 1989 concurvity is a nonlinear relationship among predictor variables which causes degeneracy of the system equations which in turn results to non-unique solutions. Moreover, in the presence of concurvity, the easy interpretability feature of additive models may no more be useful because effect of a predictor to the response variable may be affected by other variables [13, 14]. In the literature, some approaches were proposed to fit GAM models when the covariates have nonlinear relationships. Here, the question comes how the existence of linear relationship in covariates affects the model goodness of fit.

The objective of this thesis is therefore, to examine throughly the performance of three smoothing spline bases: Cubic regression spline, p-spline and thin plate spline which are commonly used in fitting GAM models under the following three situations. First, the performance of these techniques is evaluated when fitted to data containing outlier values in the response variable. In the second experiment, the methods are compared when applied to data with multicollinearity. This is to see how these GAM models perform in situations where the classic linear regression models are not appropriate due to the violence of the assumption. Finally, presence of abnormal observations in the response variable and existence of linear relationships among covariates are both considered to evaluate and compare how these GAM fitting penalized smoothing splines perform.

This thesis is organized as follows.

- **Section 2**: Discuses briefly what smoothing is and its importance. It briefly addresses splines and how they can be used to interpolate data points. It then presents how penalized splines can be used as smoothers.
- **Section 3**: This section addresses the concept of fitting GAM models as penalized GLMs. It also discusses model degrees of freedom and a method for estimating smoothing parameter.
- **Section 4**: The simulation procedures used to generate outlier-contaminated data as well as data with multicollinearity are addressed in this section.
- **Section 5**: Results and discussion of the study is presented in this section.
- **Section 6**: This final section addresses the conclusions based on the findings and includes recommendations of the thesis.

2. SMOOTHING

Smoothing is a method of estimating a nonlinear effect of one or more predictor variables on the response variable by letting the data suggest the appropriate functional form [7, 15]. There are different techniques of smoothing data. Some of the popular method of data smoothing techniques are local polynomial smoothing, kernel smoothing, regression splines, and penalized regression splines.

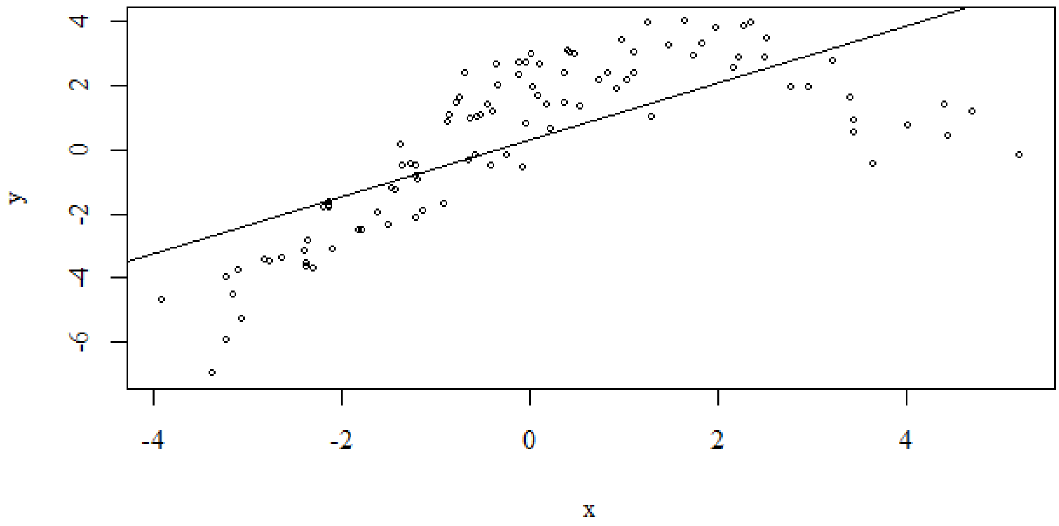


Figure 2.1. *An example where a linear model is not a best fit to the data*

In this section, the concept of splines is discussed first and later it is shown what regression splines and how to incorporate penalty to control the roughness of the curve which in turn leads to the topic of penalized regression splines are.

2.1 Splines

The term spline was originally used to name a flexible strip that was being using by draftsmen to draw curves by joining given points. The purpose was to fix the strip at its edges and calibrate it in order to pass through all the points so that the resulting shape will be used as a smooth interconnecting curve [16].

In mathematical sciences, splines are piecewise polynomial functions which are constrained to be connected at the junction points. Given a tabulated data (x_i, y_i) for

$i = 1, 2, .., n$, each point is joined by a polynomial function which results in a group of piecewise curves. A spline is then, the function which is made of these piecewise polynomials joined together at points called knots.

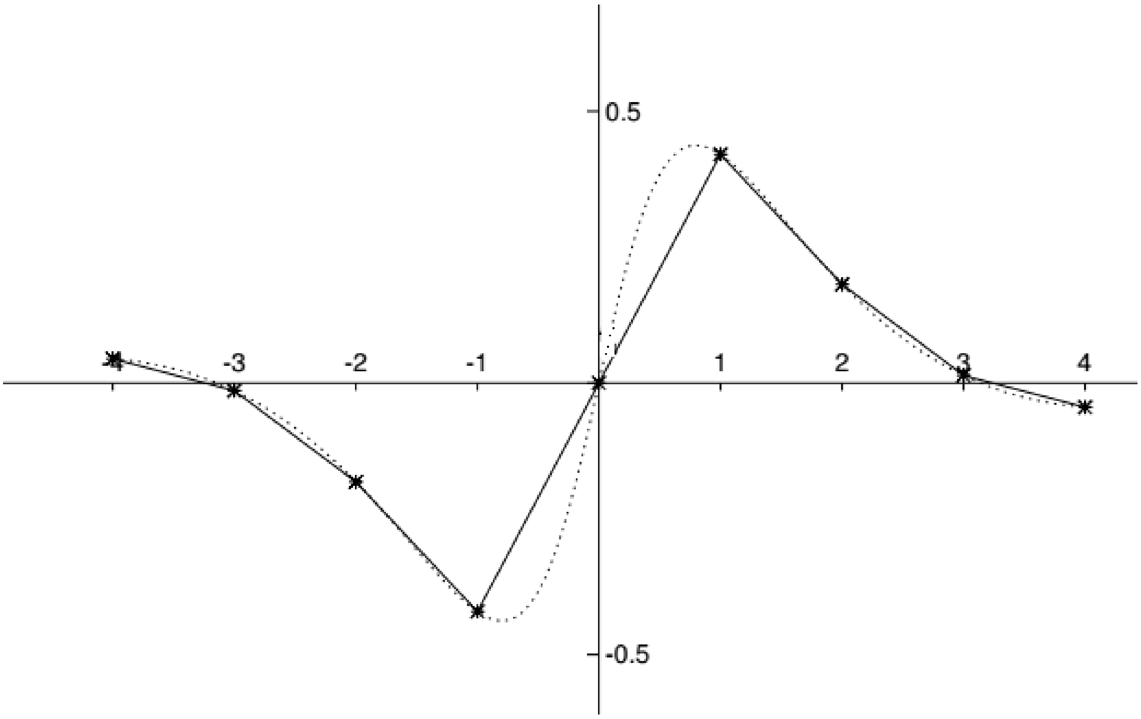


Figure 2.2. *Piecewise linear spline with 9 knots interpolating the function* $\frac{1}{1+x^2} sin(x)$ *for* $x \in [-4, 4]$.

Linear spline is a simple form of interpolation in which the piecewise functions which connect the knots are straight lines [6]. Figure 2.2 shows a simple linear spline interpolation (solid line) to estimate the function given by the dotted line. It can clearly be seen that the linear interpolation fails to capture the curvature of the function. One could use more knots to improve the accuracy of the interpolation, however, it is important to note that these kinds of interpolating functions are not continuous on their first derivatives at the knots. This can be avoided by using higher order polynomials.

Now, consider the points $\{(x_i, y_i)$, for $i = 1, 2, ..., n\}$, where, $x_i < x_{i+1}$. A function, $g(x)$, interpolating all these points which is constructed by joining sections of cubic polynomials (degree 3), one for each $[x_i, x_{i+1}]$, so that the whole function is continuous in values and on its first two derivatives is called a cubic spline [17]. Cubic splines are the most popular interpolators.

2.1.1 Natural cubic spline

Natural cubic splines are a special case of cubic splines where the second derivative at the two end points are constrained to have zero value. They are called natural for they are a solution of an optimization problem [18, 17]. In general, a natural cubic spline satisfies the following:

1. It interpolates the points (x, y) i.e. $g(x_i) = y_i$
2. Its second derivative at the two end points is zero;
$g''(x_1) = g''(x_n) = 0$.
3. Natural cubic spline is the smoothest interpolator. If $f(x)$ is any continuous function on the interval $[x_1, x_n]$ and has continuous first and second derivatives, and interpolates the points (x_i, y_i), then the natural cubic spline $g(x)$ is smoothest in the sense of minimizing the roughness measure.

$$\int_{x_1}^{x_n} g''(x)^2 dx \leq \int_{x_1}^{x_n} f''(x)^2 dx, \forall f \in C^2 \tag{2.1}$$

To demonstrate the smoothness of a cubic spline [17, 19] define a function;

$$h(x) = f(x) - g(x).$$

Since, both $f(x)$ and $g(x)$ are interpolators of the points,
$f(x_i) = g(x_i) = y_i$ which in turn leads to

$$f(x_i) = h(x_i) + g(x_i) = 0, \ \ for \ \ i = 1, 2, ..., n. \tag{2.2}$$

From Eq:2.2, $f''(x) = h''(x) + g''(x)$. Squaring both sides gives,

$$f''(x)^2 = h''(x)^2 + g''(x)^2 + 2h''(x)g''(x)$$

Integrating both sides yield,

$$\begin{aligned} \int_{x_1}^{x_n} f''(x)^2 dx = \int_{x_1}^{x_n} h''(x)^2 dx + \int_{x_1}^{x_n} g''(x)^2 dx \\ +2 \int_{x_1}^{x_n} h''(x) g''(x) dx \end{aligned} \tag{2.3}$$

Now, let the right most hand side of Eq: 2.3 is zero.

$$2 \int_{x_1}^{x_n} h''(x) g''(x) dx = 0 \tag{2.4}$$

Using the rule of integration by parts, it can be shown that

$$\big(g''(x) h'(x)\big)' = g'''(x) h'(x) + g''(x) h''(x).$$

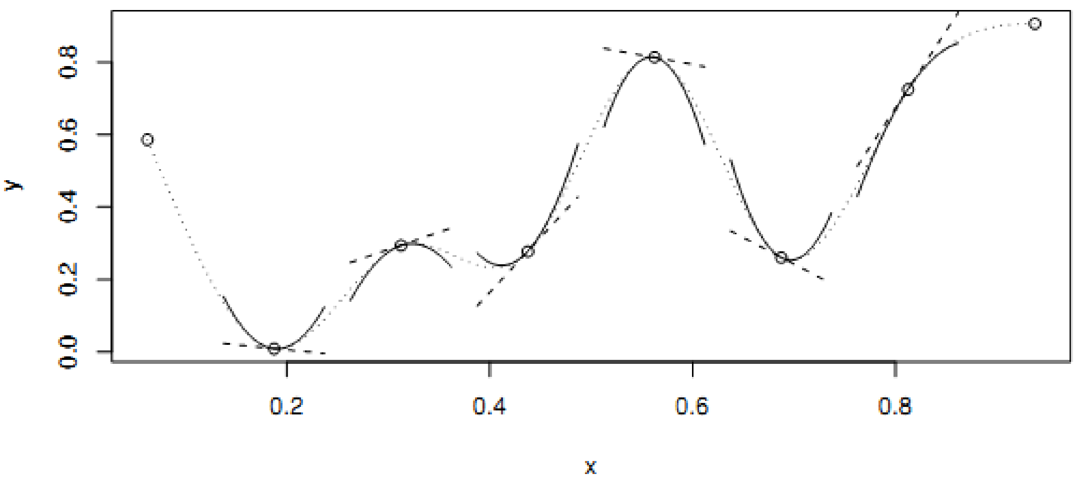


Figure 2.3. *Natural cubic spline with 7 interior knots*
Source:*Wood, 2006, p.124*

Therefore, of Eq:2.4 can be written as,

$$
\begin{aligned}
\int_{x_1}^{x_n} g''(x)h''(x)dx &= \int_{x_1}^{x_n} \left(g''(x)h'(x)\right)' dx \ - \int_{x_1}^{x_n} g'''(x)h'(x)dx \\
&= g''(x)h'(x)\bigg|_{x_1}^{x_n} - \int_{x_1}^{x_n} g'''(x)h'(x)dx \\
&= \left[g''(x_n)h'(x_n) - g''(x_1)h'(x_n)\right] - \int_{x_1}^{x_n} g'''(x)h'(x)dx \\
&= 0 - \int_{x_1}^{x_n} g'''(x)h'(x)dx
\end{aligned}
$$

By definition of natural cubic spline, since $g(x)$ is a piecewise polynomial, $g'''(x)$ is a piecewise constant, say $g'''(x) = c_i$ on each interval $[x_i, x_{i+1}]$. Furthermore, the constraints of this spline imply that

$$g''(x_1) = g''(x_n) = 0.$$

Thus,

$$
\begin{aligned}
\int_{x_1}^{x_n} g''(x)h''(x)dx &= -\sum_{i=1}^{n-1} c_i \int_{x_i}^{x_{i+1}} h'(x)dx \\
&= -\sum_{i=1}^{n-1} c_i[h(x_{i+1}) - h(x_i)]
\end{aligned}
$$

However, since both f and g are equal at the points, $h(x) = 0$ for all the knots. Therefore, as claimed in Eq:2.4,

$$\int_{x_1}^{x_n} g''(x)h''(x)dx = 0.$$

Since, the integration of the squared derivatives of f, g and h are all positive. From

Eq:2.3, it can be implied that

$$\int_{x_1}^{x_n} g''(x)^2 dx \leq \int_{x_1}^{x_n} f''(x)^2 dx = \int_{x_1}^{x_n} h''(x)^2 dx + \int_{x_1}^{x_n} g''(x)^2 dx$$

Therefore,

$$\int_{x_1}^{x_n} g''(x)^2 dx \leq \int_{x_1}^{x_n} f''(x)^2 dx, \tag{2.5}$$

where f is any twice differentiable function. The proof is complete and hence, natural cubic splines have the smoothest curvature among all smooth curves which interpolate the data points.

2.1.1.1 *Derivation of natural cubic spline*

Computationally it is tedious to find a natural cubic spline that interpolates the points (x_i, y_i), where $x_i < x_{i+1}$. However, a simple algorithm can be developed to generate a natural cubic spline $g(x)$. Let $z_i = g''(x_i)$ and $h_i = x_{i+1} - x_i$. By definition of natural cubic splines, then, $z_0 = z_n = 0$.

Lagrange form of the second derivative of the spline is given by [19],

$$g_i''(x) = \frac{z_{i+1}}{h_i}(x - x_i) - \frac{z_i}{h_i}(x - x_{i+1}).$$

By integrating $g_i''(x)$, we obtain $g_i'(x)$.

$$\int g_i''(x) dx = \int \frac{z_{i+1}}{h_i}(x - x_i) dx - \int \frac{z_i}{h_i}(x - x_{i+1}) dx$$

$$g_i'(x) = \frac{z_{i+1}}{2h_i}(x - x_i)^2 - \frac{z_i}{2h_i}(x - x_{i+1})^2 + C_i - D_i$$

Here, two arbitrary constants are added for ease of calculations.

$$\int g_i'(x) dx = \int \frac{z_{i+1}}{2h_i}(x - x_i)^2 dx - \int \frac{z_i}{2h_i}(x - x_{i+1})^2 dx + \int C_i dx - \int D_i dx$$

$$g_i(x) = \frac{z_{i+1}}{6h_i}(x - x_i)^3 - \frac{z_i}{6h_i}(x - x_{i+1})^3 + C_i(x - x_i) - D_i(x - x_{i+1})$$

Recall that cubic spline interpolates the given points (knots). Therefore,

1. $g_i(x_i) = y_i$

$$\begin{aligned} y_i =& \frac{z_{i+1}}{6h_i}(x_i - x_i)^3 - \frac{z_i}{6h_i}(x_i - x_{i+1})^3 + C_i(x_i - x_i) - D_i(x_i - x_{i+1}) \\ y_i =& - \frac{z_i}{6h_i}(-h_i)^3 - D_i(-h_i) \\ =& \frac{z_i}{6h_i}(h_i)^3 + D_i(h_i) \\ D_i =& \frac{y_i}{h_i} - \frac{z_i h_i}{6} \end{aligned} \tag{2.6}$$

2. $g_i(x_{i+1}) = y_{i+1}$

$$\begin{aligned} y_{i+1} =& \frac{z_{i+1}}{6h_i}(x_{i+1} - x_i)^3 - \frac{z_i}{6h_i}(x_{i+1} - x_{i+1})^3 + C_i(x_{i+1} - x_i) - D_i(x_{i+1} - x_{i+1}) \\ y_{i+1} =& \frac{z_{i+1}}{6h_i}(h_i)^3 + C_i(h_i) \\ C_i =& \frac{y_{i+1}}{h_i} - \frac{z_{i+1} h_i}{6} \end{aligned} \tag{2.7}$$

Substituting C_i and D_i in $g(x)$ yields

$$g_i(x) = \frac{z_{i+1}}{2h_i}\big(x - x_i\big)^3 - \frac{z_i}{6h_i}\big(x - x_{i+1}\big)^3 + \big(\frac{y_{i+1}}{h_i} - \frac{h_i z_{i+1}}{6}\big)\big(x - x_i\big) - \big(\frac{y_i}{h_i} - \frac{z_i h_i}{6}\big)\big(x - x_{i+1}\big) \tag{2.8}$$

and first derivative of $g(x)$ is given by,

$$g_i'(x) = \frac{z_{i+1}}{2h_i}(x - x_i)^2 - \frac{z_i}{2h_i}(x - x_{i+1})^2 + \frac{y_{i+1} - y_i}{h_i} - \frac{z_{i+1} - z_i}{6} h_i$$

Let the fraction $\frac{y_{i+1} - y_i}{h_i}$ which is a constant be denoted by b_i. By definition of cubic splines, $g(x)$ has continuous first derivative;

$$\begin{aligned} g_{i-1}'(x_i) &= g_i'(x_i), \quad \text{for} \quad i = 1, 2, ..., n-1. \\ g_i'(x_i) &= -\frac{h_i z_i}{2} + b_i - \frac{z_{i+1} - z_i}{6} h_i \end{aligned}$$

and,

$$\begin{aligned} g_{i-1}'(x_i) &= \frac{z_i}{h_i}(x - x_{i-1})^2 - \frac{z_{i-1}}{2h_{i-1}}(x - x_i)^2 + b_{i-1} - \frac{z_i - z_{i-1}}{6} h_{i-1} \\ &= \frac{z_i}{2} h_{i-1} + b_{i-1} - \frac{z_i - z_{i-1}}{6} h_{i-1} \end{aligned}$$

Equating $g'_{i-1}(x_i)$ and $g'_i(x_i)$,

$$\begin{aligned}
g'_{i-1}(x_i) - g'_i(x_i) =& 0 \\
\frac{z_i h_{i-1}}{2} - \frac{z_i - z_{i-1}}{6} h_{i-1} + \frac{h_i z_i}{2} + \frac{z_{i+1} - z_i}{6} h_i =& b_i - b_{i-1} \\
3(z_i h_{i-1}) - (z_i - z_{i-1}) h_{i-1} + 3 h_i z_i + (z_{i+1} - z_i) h_i =& 6(b_i - b_{i-1}) \\
h_{i-1} z_{i-1} + 2(h_{i-1} + h_i) z_i + h_i z_{i+1} =& 6(b_i - b_{i-1})
\end{aligned}$$

There are a total of $n - 1$ linear equations. Since $z_0 = z_n = 0$, the first and last equations have only two terms at the left hand side of the equality. The linear equations can be written in a matrix form;

$$H.z = b \tag{2.9}$$

where,

$$H = \begin{pmatrix}
2d_1 & h_1 & 0 & 0 & \cdots & 0 & 0 & 0 \\
h_1 & 2d_2 & h_2 & 0 & \cdots & 0 & 0 & 0 \\
0 & h_2 & 2d_3 & h_3 & \cdots & 0 & 0 & 0 \\
0 & 0 & h_3 & 2d_4 & \ddots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 2d_{n-3} & h_{n-3} & 0 \\
0 & 0 & 0 & 0 & \cdots & h_{n-3} & 2d_{n-2} & h_{n-2} \\
0 & 0 & 0 & 0 & \cdots & 0 & h_{n-2} & 2d_{n-1}
\end{pmatrix}$$

$$z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_{n-2} \\ z_{n-1} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 6(b_1 - b_0) \\ 6(b_2 - b_1) \\ 6(b_3 - b_2) \\ \vdots \\ 6(b_{n-2} - b_{n-3}) \\ 6(b_{n-1} - b_{n-2}) \end{pmatrix}$$

where $d_i = h_{i-1} + h_i$, $h_i = x_{i+1} - x_i$ and $b_i = \frac{y_{i+1} - y_i}{h_i}$.

The fact that H is symmetric, tridiagonal, and diagonal dominant ensures that Eq:2.9 has unique solutions for the unknown variables ($z'_i s$). Once obtained, substituting values of z_i's in Eq:2.8 will result in a cubic polynomial function which is the desired natural cubic spline.

2.1.2 Roughness measure of a curve

There are different ways of measuring roughness of a curve g defined on an interval $[a, b]$ [17]. Given it is twice differentiable function, one common way of quantifying the wiggliness of g is given by the integration of its squared second derivative;

$$J(g) = \int_a^b \left[g''(x)\right]^2 dx \tag{2.10}$$

Some motivating ways of this measure of wiggliness are explained in Green and Silverman [17]. From the mathematical perspective, $|g''(x)|$ measures the turning rate of the curve at a specific value of x [20]; this measure should not be affected by a linear or constant terms in the model. Thus, two functions which differ only by a constant or a linear term should have the same quantity of second derivatives at a given point. Based on this intuitive idea, it becomes logical to use Eq:2.10 as a measure of global wiggliness of a curve.

2.2 Smoothing Bases

A smoothing function can best be described by considering a univariate model with only one smoothing function.

$$y_i = f(x_i) + \epsilon_i \tag{2.11}$$

where, f is a smoother, and ϵ_i are i.i.d random variables which follows a distribution of $N(0, \sigma^2)$.

Here, a basis expansion can be used to define the dimensions of the model matrix so that a more flexible smoothing function can be achieved [21]. Choosing the basis function allows the smoother to be written as a linear combination of these bases.

Let $b_i(x)$ be the i^{th} basis function for $i = 1, 2, ..., q$. Then, f is written as;

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i \tag{2.12}$$

where, β_i's are unknown parameters. Therefore, substituting Eq:2.12 in Eq:2.11 yields a linear model given by,

$$y_i = b_1(x)\beta_1 + b_2(x)\beta_2 + \ldots + b_q(x)\beta_q + \epsilon_i \tag{2.13}$$

Different functions can be employed to define the bases. Some common bases are discussed below.

2.2.1 Polynomial basis

Based on how many basis functions should be used to estimate the smoother, a given order polynomial would be chosen. For example, if a fourth order is chosen, the bases functions are given by, $b_1(x) = 1$, $b_2(x) = x$, $b_3 = x^2$, $b_4 = x^3$ and $b_5 = x^4$ and the smoothing function is then written as,

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5. \tag{2.14}$$

The i^{th} row of the model matrix is given by,

$$\boldsymbol{X}_i = \left[1, x_i, x_i^2, x_i^3, x_i^4\right].$$

Here, an ordinary least square method is used to determine the model parameters.

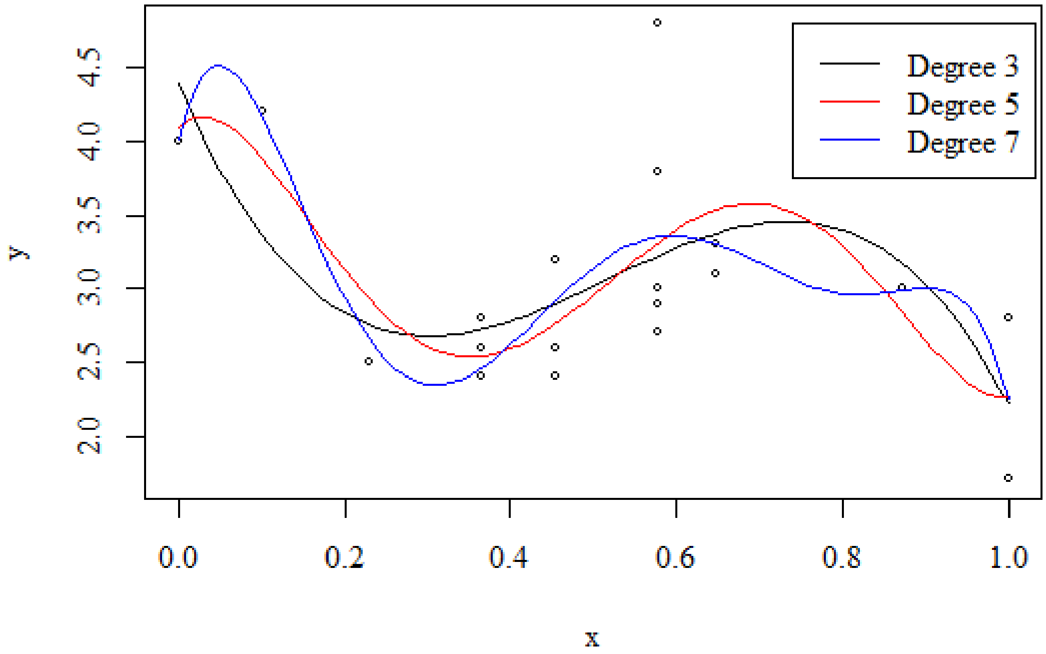


Figure 2.4. *An illustration of smoothing data using polynomial bases of varying degrees.*

In polynomial regression approach, a higher-order may generally represent the data well [18]. However, using a higher-degree polynomial means more model parameters will be included which may result in over-fitting. Figure 2.4 shows how an increase in order of the polynomial basis influences goodness of fit of the model.

2.2.2 Spline bases

Splines are among the popular ways of smoothing data by fitting the underlying function. They are more advantageous than their polynomial counterparts for different reasons [6].

1. Splines are superior in a sense that they have analytical foundation; it can be proved that a spline smoother provides a fit with a minimum mean square error.
2. In smoothing splines, a term which controls the trade-off between over-fitting and goodness of fit can be added to the optimization problem which polynomial regression lacks to have.
3. For the reason that a lot of new studies about splines are being done while that of polynomials are more or less static, softwares which implement splines are superior to those which implement polynomial regression.

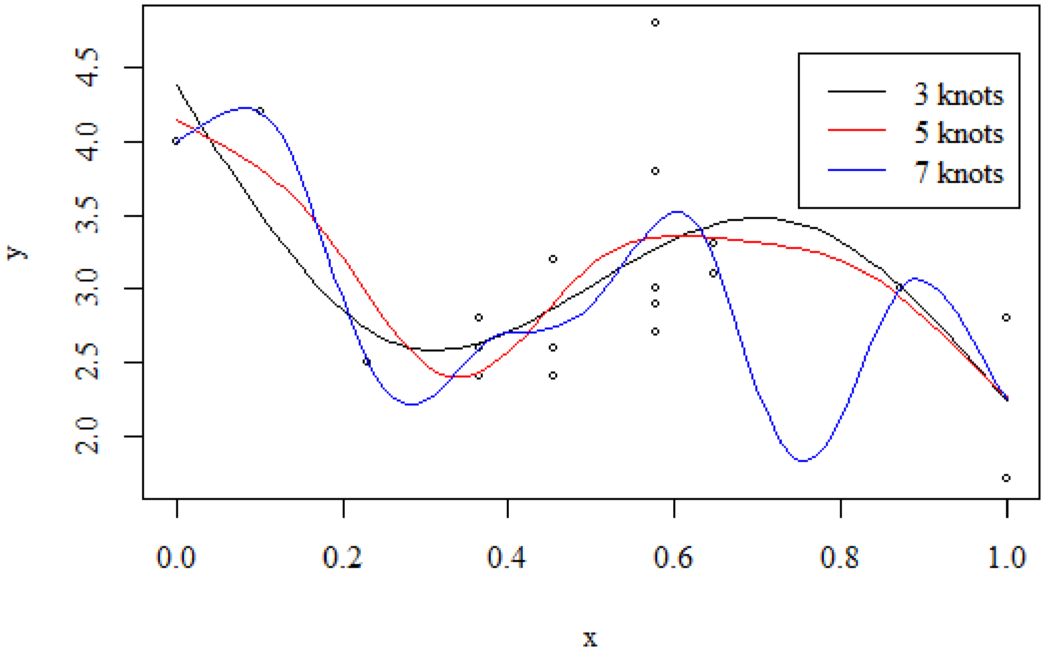


Figure 2.5. *Cubic regression spline with different number of knots.*

As discussed before, (natural) cubic splines are the smoothest among all interpolators. For this reason, a cubic regression spline will be considered to show how splines are used to fit a univariate model. Given knots, x_j^* for $j = 1, 2, ..., q$, there are different ways of representing cubic basis. For simplicity [4, 22], consider the following simple representation of cubic smoothing spline for $x \in [0, 1]$. The cubic bases are given by, $b_1(x) = 1$, $b_2(x) = x$, and $b_{j+2}(x) = R(x, x_j^*)$, where,

$$R(x, x_j^*) = \frac{\left[(x_j^* - \frac{1}{2})^2 - \frac{1}{12}\right]\left[(x - \frac{1}{2})^2 - \frac{1}{12}\right]}{4} - \frac{\left[(|x - x_j^*| - \frac{1}{2})^4 - \frac{1}{2}(|x - x_j^*| - \frac{1}{2})^2 + \frac{7}{240}\right]}{24} \tag{2.15}$$

The smoother is then given by,

$$f(x) = \beta_1 + \beta_2 x + \sum_{j=1}^{q} \beta_{j+2} R_i(x, x_j^*) \tag{2.16}$$

The model matrix is an n by $q + 2$ where, the i^{th} row is given by,

$$\boldsymbol{X}_i = \left[1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), ..., R(x_i, x_q^*)\right] \tag{2.17}$$

In Figure 2.5, the model does not seem good enough to fit the data when three knots are chosen. On the other hand, if too many knots are used (e.g. the fit with seven knots) the problem of over-fitting occurs; the fit becomes too wiggly. Therefore, one has to choose appropriate number of knots in order to get the best fit. In reality, this is somehow subjective and not the best way to smooth large data for it is hard to guess the appropriate number of knots. A wiggliness controlling mechanism should then be addressed in modeling with smoothing functions [4].

As discussed above, the model roughness is controlled by setting the basis dimension appropriately fixed. It can clearly be seen that the goodness of fit of the model depends on the location and number of knots. Using too many knots results too wiggly fit and on the other hand, a model with fewer knots may not fit the data well.

A better way of controlling the wiggliness of a model fit is by using a penalized regression spline where a roughness of penalty is added to the least square fitting optimization in order to control the smoothness. In this case, the number of knots are chosen to be a little more than believed could be desired [4]. The objective in penalized regression spline is to fit the model by minimizing,

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \int \left[f''(x)\right]^2 dx \tag{2.18}$$

where, the integrated term is the roughness measure, λ is a smoothing parameter which controls the trade-off between smoothness and goodness of fit of a model. If $\lambda \to \infty$, f

will approach to a straight line fit, where as if $\lambda = 0$, f will be identical to the regression spline fitting technique for it will not be penalized.

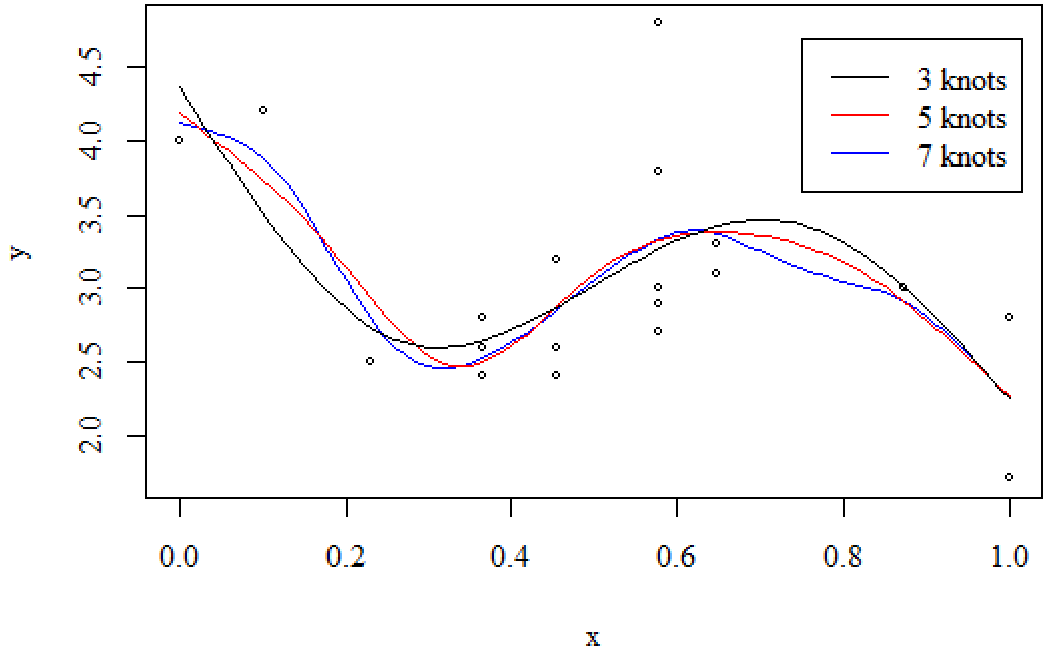


Figure 2.6. *Penalized regression spline with different number of knots*

Wood, 2006 showed that the penalty term can be written in quadratic form of $\boldsymbol{\beta}$;

$$\int \left[f''(x)\right]^2 dx = \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S} \boldsymbol{\beta} \tag{2.19}$$

Replacing Eq:2.19 in Eq:2.18 yields,

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S} \boldsymbol{\beta} \tag{2.20}$$

Computationally, Eq:2.20 is more favorable in comparison with the somewhat complicated form of Eq:2.18. Elements of matrix $\boldsymbol{S}$ are known coefficients which are calculated using the given knots by using Eq:2.15 and is written as,

$$\boldsymbol{S} = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & R(x_1^*, x_1^*) & R(x_1^*, x_2^*) & \cdots & R(x_1^*, x_{q-1}^*) & R(x_1^*, x_q^*) \\ 0 & 0 & R(x_2^*, x_1^*) & R(x_2^*, x_2^*) & \cdots & R(x_2^*, x_{q-1}^*) & R(x_2^*, x_q^*) \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & R(x_q^*, x_1^*) & R(x_q^*, x_2^*) & \cdots & R(x_q^*, x_{q-1}^*) & R(x_q^*, x_q^*) \end{pmatrix}$$

In a similar way to that of ordinary least square method, the penalized least square estimator, minimizer of Eq:2.18 is

$$\begin{aligned} \frac{\partial}{\partial \beta}\left[(y - \boldsymbol{X\beta})^T(y - \boldsymbol{X\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{S\beta}\right] &= 0 \\ (y - \boldsymbol{X\beta})(-\boldsymbol{X}^T) + \lambda \boldsymbol{S\beta} &= 0 \\ \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \end{aligned}$$

and the hat or influence matrix is then given by,
$\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1}\boldsymbol{X}^T$.
Further details of this estimator and its implementations can be found in Wood, 2006.

In the previous section (see Figure 2.5), it has been clearly shown how the increase of knot size disturbs the model smoothness. Moreover, Figure 2.6 shows how penalized regression method controls the roughness of the model by adding a penalty term to the minimization objective. As it can be seen from those two graphs, the more wiggly 7-knots model in the unpenalized regression can safely be argued that it produced the best fit after introducing the penalty term.

Furthermore, Figure 2.7 illustrates smoothing data with polynomial basis of order 4, cubic regression spline (unpenalized) and penalized (cubic) regression spline. The penalized fit seems to represent the data better. Therefore, it is reasonable to adopt penalized regression spline for smoothing.

Following, three smoothing spline techniques which incorporate a penalty term will be discussed. These are among the commonly used smoothing splines which are implemented in the *mgcv* package [4].

2.2.2.1 *Cubic spline*

There are different possible ways of defining a cubic smoothing spline basis. One simple representation of cubic spline was introduced in previous sections. Another way of

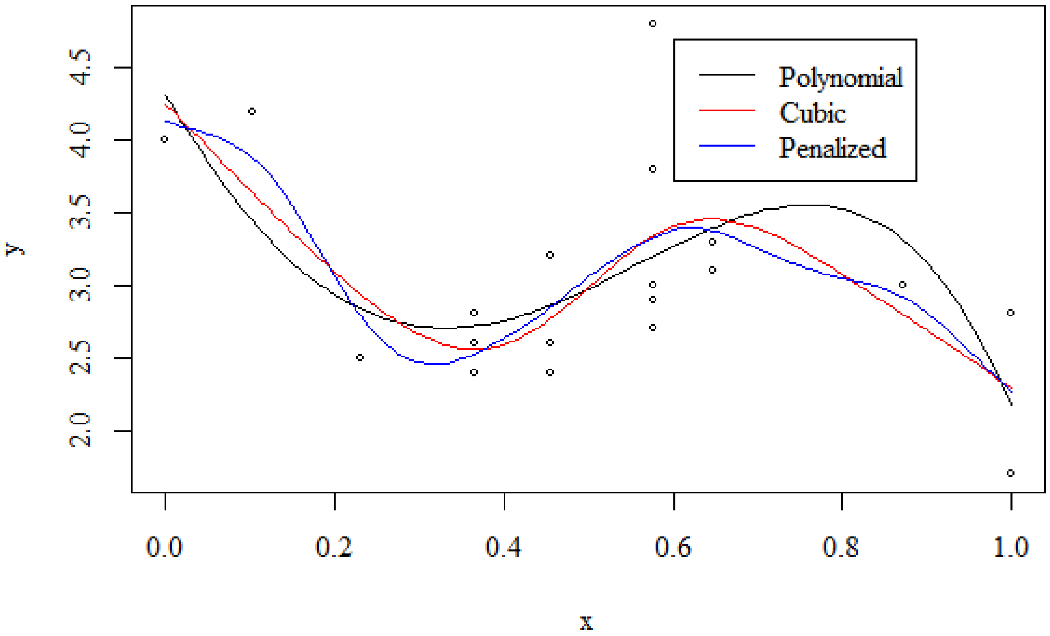


Figure 2.7. *Representing univariate smoothing functions using different bases; Polynomial basis of degree 4, cubic regression spline with 4 knots, and penalized regression spline with 7 knots.*

representing this kind of splines which has an advantage of easy interpretation of model parameters [4] is given by Eq:2.21. In this approach, the parameters are given in terms of the knot values.

Recall constructing a cubic spline function $f(x)$, defined for knots $x_1, x_2, ..., x_k$. If we let $\beta_j = f(x_j)$ and $\delta_j = f''(x_j)$, then $f(x)$ with basis functions a_j^-, a_j^+, c_j^-, and c_j^+ is given by,

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1} \quad \text{if} \quad x_j \leq x \leq x_{j+1} \tag{2.21}$$

The basis functions are defined in Table 2.1. It can easily be shown that $\beta_j = f(x_j)$ and $\delta_j = f''(x_j)$.

Table 2.1. *Basis functions for a cubic spline;* $h_j = x_{j+1} - x_j$

$a_j^-(x) = \frac{x_{j+1}-x}{h_j}$	$c_j^-(x) = \frac{\frac{(x_{j+1}-x)^3}{h_j} - h_j(x_{j+1}-x)}{6}$
$a_j^+(x) = \frac{x-x_j}{h_j}$	$c_j^+(x) = \frac{\frac{(x-x_j)^3}{h_j} - h_j(x-x_j)}{6}$

Re-writing $f(x)$ by putting the basis functions in,

$$f(x) = \frac{x_{j+1} - x}{h_j}\beta_j + \frac{x - x_j}{h_j}\beta_{j+1} + \frac{\frac{(x_{j+1}-x)^3}{h_j} - h_j(x_{j+1} - x)}{6}\delta_j + \frac{\frac{(x-x_j)^3}{h_j} - h_j(x - x_j)}{6}\delta_{j+1} \tag{2.22}$$

To find $f(x_j)$, substitute x_j in Eq:2.22

$$\begin{aligned} f(x_j) &= \frac{x_{j+1} - x_j}{h_j}\beta_j + \frac{x_j - x_j}{h_j}\beta_{j+1} + \frac{\frac{(x_{j+1}-x_j)^3}{h_j} - h_j(x_{j+1} - x_j)}{6}\delta_j \\ &\qquad + \frac{\frac{(x_j-x_j)^3}{h_j} - h_j(x_j - x_j)}{6}\delta_{j+1} \\ &= \frac{h_j}{h_j}\beta_j + 0 + \frac{h_j^2 - h_j^2}{6}\delta_j + 0 \\ &= \beta_j \end{aligned}$$

The second derivative of $f(x)$ is given by,

$$f''(x) = \frac{x_{j+1} - x}{h_j}\delta_j + \frac{x - x_j}{h_j}\delta_{j+1}$$

Therefore, similarly substituting x_j in $f''(x_j)$ yields,

$$\begin{aligned} f''(x_j) &= \frac{x_{j+1} - x_j}{h_j}\delta_j + \frac{x_j - x_j}{h_j}\delta_{j+1} \\ &= \frac{h_j}{h_j}\delta_j + 0 \\ &= \delta_j \end{aligned}$$

By definition, a natural cubic spline is continuous to second derivative and has zero second derivative values at the two end knots, say x_1 and x_k. In this case, to show that $f(x)$ satisfies these properties is equivalent to showing Eq:2.23 holds true.

$$\boldsymbol{B}\boldsymbol{\delta}^- = \boldsymbol{D}\boldsymbol{\beta} \tag{2.23}$$

where, $\boldsymbol{\delta}^- = (\delta_2, \delta_3, ..., \delta_{k-1})^T$ $\left[\text{because } f''(x_1) = \delta_1 = \delta_k = f''(x_k) = 0\right]$. The matrices $\boldsymbol{B}$ and $\boldsymbol{D}$ are given below. Continuity of the first derivative of the spline implies that the the derivative of the sections to the left and right x_j are equal [4]. Hence, it can be written as;

$$-\frac{\beta_j}{h_j}+\frac{\beta_{j+1}}{h_j}+\delta_j\frac{h_j}{6}+\delta_{j+1}\frac{3h_j}{6}-\delta_{j+1}\frac{h_j}{6}=-\frac{\beta_{j+1}}{h_{j+1}}+\frac{\beta_{j+2}}{h_{j+1}}-\delta_{j+1}\frac{3h_{j+1}}{6}+\delta_{j+1}\frac{h_{j+1}}{6}-\delta_{j+2}\frac{h_{j+1}}{6}$$

Multiplying both sides by -1 and a simple rearrangement of the above equality results in,

$$\frac{1}{h_j}\beta_j+\left(\frac{1}{h_j}+\frac{1}{h_{j+1}}\right)\beta_{j+1}+\frac{1}{h_{j+1}}\beta_{j+2}=\frac{h_j}{6}\delta_j+\big(\frac{h_j}{3}+\frac{h_{j+1}}{3}\big)\delta_{j+1}+\frac{h_{j+1}}{6}\delta_{j+2}$$

For $j=1,2,...,k-2$, the terms in Eq:2.23 can be written as;

$$\boldsymbol{D}=\begin{pmatrix} \frac{1}{h_1} & (\frac{1}{h_1}+\frac{1}{h_2}) & \frac{1}{h_2} & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & h_2 & (\frac{1}{h_2}+\frac{1}{h_3}) & \frac{1}{h_3} & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & h_3 & (\frac{1}{h_3}+\frac{1}{h_4}) & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{h_{k-2}} & (\frac{1}{h_{k-2}}+\frac{1}{h_{k-1}}) & \frac{1}{h_{k-1}} \end{pmatrix}$$

and

$$\boldsymbol{B}=\begin{pmatrix} 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \frac{h_1}{6} & \left(\frac{h_1}{3}+\frac{h_2}{3}\right) & \frac{h_2}{6} & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \frac{h_2}{6} & \left(\frac{h_2}{3}+\frac{h_3}{3}\right) & \frac{h_3}{6} & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{h_3}{6} & \left(\frac{h_3}{3}+\frac{h_4}{3}\right) & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{h_{k-2}}{6} & \left(\frac{h_{k-2}}{3}+\frac{h_{k-1}}{3}\right) & \frac{h_{k-1}}{6} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \end{pmatrix}$$

In Eq:2.23, let $\boldsymbol{F}^{-}=\boldsymbol{B}^{-1}\boldsymbol{D}$, and by augmenting this new matrix as,

$$\boldsymbol{F}=\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{F}^{-} \\ \boldsymbol{0} \end{bmatrix}$$

$\boldsymbol{\delta}$ can be written in terms of $\boldsymbol{B}$, $\boldsymbol{D}$ and $\boldsymbol{\beta}$ as, $\boldsymbol{\delta}=\boldsymbol{F\beta}$. Consequently, we can write Eq:2.21 in terms of $\boldsymbol{\beta}$ as,

$$f(x)=a_j^{-}(x)\beta_j+a_j^{+}(x)\beta_{j+1}+c_j^{-}(x)\boldsymbol{F}_j\boldsymbol{\beta}+c_j^{+}(x)\boldsymbol{F}_{j+1}\boldsymbol{\beta},\ x_j\le x\le x_{j+1} \tag{2.24}$$

and this can be re-written in the general form as

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i$$

The wiggliness measure of the spline [4] is given by,

$$\int_{x_1}^{x_k} (f''(x))^2 dx = \boldsymbol{\beta^T D^T B^{-1} D\beta} \tag{2.25}$$

where the penalty matrix is $\boldsymbol{S} = \boldsymbol{D^T B^{-1} D}$.

2.2.2.2 *P-spline*

For (natural) cubic spline, there is high tendency for the columns of the model matrix $\boldsymbol{X}$, to be correlated for they are in someway transformed version of the predictor variable(s) [6]. This dependency may cause multicollinearity or concurvity which may result in numerical instability and imprecision in the spline fit [6, 18]. To somehow get rid off this problem, a *B-spline* basis which is refined form of a cubic spline, can be employed. This kind of splines can be used to represent cubic splines as well as higher order splines.

B-spline basis, a strictly local type of spline is non-zero only on the intervals between $m + 3$ adjacent knots where $m + 1$ is the order of the basis (i.e. $m = 2$ for cubic spline) [4]. In B-spline basis, $m + 1$ knots are added on two sides of the specified knots so that totally there will be $(m + 1) + k + (m + 1)$ knots. The spline is however, defined only on the interval $[x_{m+2}, x_k]$ which implies that the first $m + 1$ and last $m + 1$ knots are arbitrary. Any spline of order $m + 1$ can then be represented as:

$$f(x) = \sum_{i=1}^{k} B_i^m(x)\beta_i \tag{2.26}$$

where the B-spines can recursively be written as,

$$B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x), \quad \text{i} = 1, 2, ..., k$$

Therefore, based on Eq:2.26 a cubic B-spline function ($m = 2$) with its B-spline bases are respectively written as;

$$f(x) = \sum_{i=1}^{k} B_i^2(x)\beta_i \tag{2.27}$$

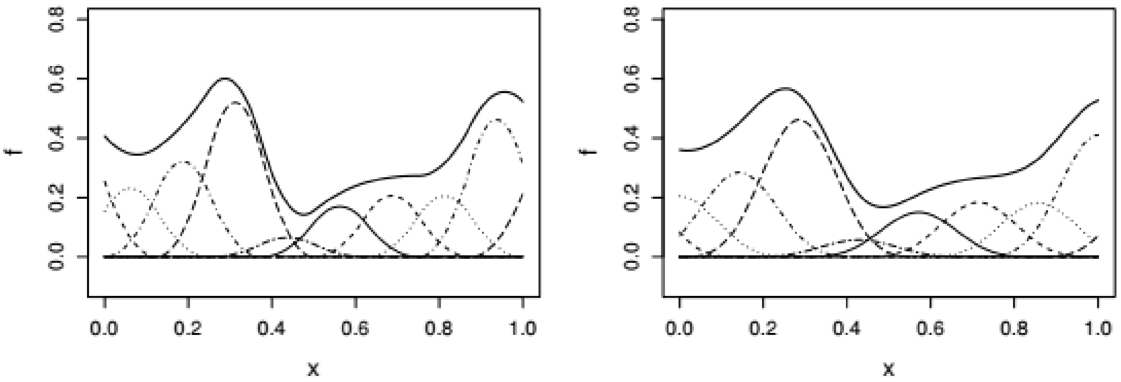


Figure 2.8. *An illustration of representing a smooth curve by B-spline. Dashed (or dotted) curves are B-spline bases functions multiplied by their coefficients where each nonzero over 3 intervals in the left panel (m=1), and 4 intervals in the right panel (m=2) (solid curves represents the desired curve).*
Source: *Wood, 2006, p.153*

and each of the B-spline bases (B's) for $i = 1, 2, \ldots, k$ are given by,

$$\begin{aligned}
B_i^2(x) &= \frac{x - x_i}{x_{i+3} - x_i} B_i^1(x) + \frac{x_{i+4} - x}{x_{i+4} - x_{i+1}} B_{i+1}^1(x) \\
B_i^1(x) &= \frac{x - x_i}{x_{i+2} - x_i} B_i^0(x) + \frac{x_{i+3} - x}{x_{i+3} - x_{i+1}} B_{i+1}^0(x) \\
B_i^0(x) &= \frac{x - x_i}{x_{i+1} - x_i} B_i^{-1}(x) + \frac{x_{i+2} - x}{x_{i+2} - x_{i+1}} B_{i+1}^{-1}(x) \\
B_i^{-1}(x) &= \begin{cases} 1 & \text{if } x_i \le x < x_{i+1} \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{2.28}$$

Basically, a B-spline basis is a rescaling of the piecewise functions, which is the same principle with rescaling explanatory variables by mean subtraction in order to minimize collinearity [6]. In a very similar way, the rescaling in B-spline reduces collinearity between the bases of the model matrix $\boldsymbol{X}$. This is generally true if large number of knots are used, otherwise, the B-spline would not be stable [4, 6, 23].

P-splines, a penalty incorporated B-splines, are proposed by Eilers and Marx, 1996 as a more stable version of the B-spline bases particularly for lower rank smoothing. They are generally defined on an equidistant knots and use a difference penalty applied to adjacent coefficients, β_i, directly. For example, as given in Wood, 2006 if a squared difference of adjacent parameters is to be used as penalty measure, it looks like,

$$\mathcal{P} = \sum_{i=1}^{k-1} (\beta_{i+1} - \beta_i)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + 2\beta_3^2 + \ldots + \beta_k^2 \tag{2.29}$$

In matrix form, it can be written as,

$$\mathcal{P} = \beta^T \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix} \beta$$

By increasing the differences parameter, a higher order penalty can be produced. The advantage of p-spline is that they are easy to set up and use. Additionally, they are flexible in the sense that any order of penalty can be incorporated to any order of B-spline basis. However, if knots are unevenly spaced, p-splines lose their simplicity behavior [4].

2.2.2.3 *Thin plate spline*

For an arbitrary spaced data (x_i,y_i), thin plate spline, say $f(x_i, y_i)$, is a two-dimensional interpolation scheme which is an extension of the natural cubic spline for one dimensional data [25]. Splines of these types are good solutions for the smoothing function problem of more than one predictor variables [4]. In thin plate spline, the problem of estimating a smoothing function, f, is an estimation of a surface while in natural cubic spline, it is a curve estimation problem [17].

Green and Silverman, 1994 put forth the general properties of the extended cubic spline in order to develop a methodology (thin- plate interpolant) for bivariate (for simplicity a bivariate case is used) case. The properties of the roughness penalty, say J, for data points (x_1, x_2) can be summarized as,

1. If the second derivatives of f are square-integrable over $\Re^2$, J is finite.

2. If f has high local curvature, J will be large resulting in a large second derivative. Intuitively, it can be seen that J measures the wiggliness of f.

3. Rotating the coordinates in $\Re^2$ does not affect J.

4. The wiggliness penalty, J, is zero if and only if f is a linear function.

In smoothing procedure, J is used as roughness penalty and in interpolation, subjected to the interpolation conditions, it is used to find the natural thin-plate interpolator [17].

Now, consider the smoothing function, $f(\boldsymbol{x})$, estimation problem,

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i$$

where ϵ_i is the random error term and $\boldsymbol{x}$ is a d−vector from n ($\geq d$) observations (x_i, y_i). In this case , thin plate can be used to estimate the smoothing function, f, of the data points $(\boldsymbol{x}_i, y_i)$, where $i = 1, 2, ..., n$ ($n \geq d$) by finding the function g [4, 26] which minimizes

$$\|\boldsymbol{y} - \boldsymbol{g}\|^2 + \lambda J_{md}(g) \tag{2.30}$$

where, $\boldsymbol{g}$=$(g(x_1), g(x_2), ..., g(x_n))^T$, $\boldsymbol{y}$=$(y_1, y_2, ..., y_n)^T$. $J_{md}(g)$ is the penalty function which measures the wiggliness of the smoother, g, whereas, λ is the smoothing parameter. Here, the roughness penalty function [4, 17, 26] is given by

$$J_{md} = \int \cdots \int_{\Re^d} \sum_{v_1+v_2+...+v_d=m} \frac{m!}{v_1!...v_d!} \left(\frac{\partial^m g}{\partial x_1^{v_1} ... \partial x_d^{v_d}} \right)^2 dx_1...dx_d. \tag{2.31}$$

For two dimension (d = 2, m = 2), this measure of wiggliness is written as

$$J_{22} = \int \int \left(\frac{\partial^2 g}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 g}{\partial x_2^2} \right)^2 dx_1 dx_2$$

Given the restriction $2m > d$ is true, Wood, 2006 put forth that the minimizer of Eq:2.30 has the form of

$$g(x) = \sum_{n=1}^{n} \delta_i \eta_{md}(||x - x_i||) + \sum_{j=1}^{m} \alpha_j \phi_j(x), \tag{2.32}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are vectors of coefficients to be estimated in which $\boldsymbol{\delta}$ is subject to the linear constraint $\boldsymbol{T}^T \boldsymbol{\delta} = \boldsymbol{0}$ where $T_{ij} = \phi_j(x_i)$. The ϕ_j functions span 'null space' of functions for which J_{md} is zero. If $m = d = 2$ for example, these basis functions are given by $\phi_1(\boldsymbol{x}) = 1$, $\phi_2(\boldsymbol{x}) = x_1$, and $\phi_3(\boldsymbol{x}) = x_2$. Moreover, the other basis functions in Eq:2.32 are given by

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d} log(r) & d \text{ is even} \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & d \text{ is odd} \end{cases} \tag{2.33}$$

Here, it is important to note that thin plate splines can be used for any number of predictors [4]. In addition, there is no need of specifying knot positions. On the other hand, the disadvantage of these kind of smoothers is their being computational expensive; there are as many parameters to be estimated as there are data points [26].

2.2.3 Additive models

Now let us consider a model with two covariates, x and z, for a response variable y_i. Then Eq:2.11 is extended in an additive form as,

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i \tag{2.34}$$

f_j's are smoothers, and ϵ_i are i.i.d random variables $N(0, \sigma^2)$. From Eq:2.34, it can be noted that the effects of the covariates to the response variable are assumed to be strictly additive. Additive models (AM) as well can be represented using penalized regression spline in a similar way to that of univariate models. Each smoother can independently be written as a linear combination of the basis.

$$f_1(x) = \delta_1 + \delta_2 x + \sum_{i=1}^{q_1} \delta_{i+2} R_i(x, x_i^*)$$

and

$$f_2(x) = \gamma_1 + \gamma_2 x + \sum_{i=1}^{q_2} \gamma_{i+2} R_i(z, z_i^*)$$

where f_1 has $q_1 + 2$ unknown parameters, δ_i, whereas f_2 has $q_2 + 2$ parameters, γ_i. The knot locations of these two smoothers are given by x_i^* and z_i^* respectively.

The problem of identifiability in additive models can clearly be seen from these equations where the constants are confounded. This can easily be avoided by constraining one of them to zero [4]. Let $\gamma_i = 0$, then the i^{th} row of the additive model matrix is

$$\boldsymbol{X}_i = [1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), ..., R(x_i, x_{q1}^*), z_i, R(z_i, z_1^*), R(z_i, z_2^*), ..., R(z_i, z_{q2}^*)]$$

and similarly, the conjugated parameters of the two smoothers will be,

$$\boldsymbol{\beta} = (\delta_1, \delta_2, ..., \delta_{q1}, \gamma_2, \gamma_3, ..., \gamma_{q2}).$$

Having all this, the roughness measure of the smoothers can then be written [4] in exactly similar way to Eq:2.10 as,

$$\int f_1''(x)^2 dx = \boldsymbol{\beta^T S_1 \beta} \quad \text{and} \quad \int f_2''(x)^2 dx = \boldsymbol{\beta^T S_2 \beta}$$

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

The optimization problem to be minimized in order to fit the additive model using penalized least squares method is then given by,

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

The smoothing parameters, λ_1 and λ_2, control the smoothness of f_1 and f_2 respectively and give more weight to the one which is more close to the objective model.

Similarly, the additive model with two covariates discussed here can be extended into a model with more covariates. Moreover, bases other than the above given cubic regression basis can be used to fit additive models.

3. GENERALIZED ADDITIVE MODELS

In this chapter, the nonparametric (or semi-parametric) regression model, generalized additive models (GAMs), will be discussed. The core point of this chapter will be to present a commonly used method of fitting a GAM model; the smoothing spline basis approach which is used in *mgcv* package [28]. Given a spline basis is selected, a GAM model fitting will be discussed in relation to that of GLMs. For this reason, first, generalized linear (GLM) model and its fitting method will be discussed.

In application, fitting GAM is an estimation of model parameters as well as smoothing parameters. Appropriate techniques of estimating smoothing parameters will be presented as well as a part of the model estimation. Furthermore, multicollinearity and outliers in a data and their effect to a fitted model will be addressed.

3.1 Generalized Linear Models (GLMs)

Recall that generalized linear models (GLMs) are an extension of the ordinary linear regression model [3] in a sense that they can model any response variable which follows any of the exponential family distributions. GLMs consider a response variable y whose distribution is from the exponential family distributions [2, 4, 27]. Any distribution with a probability density function of the form

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\right] + h(y_i,\phi_i) \tag{3.1}$$

belongs to the exponential family distributions. Here, a, b, and h are any arbitrary functions, ϕ is a scale parameter, θ_i is the so called natural location parameter, whereas y_i for $i = 1, 2, ..., n$ represents the predictor variables. For exponential family members [2],

$$\begin{aligned} \mu = E(y) &= b'(\theta) \\ V(y) &= b''(\theta)a(\phi) \end{aligned} \tag{3.2}$$

Here, $a(\phi)$ is any function of the scale parameter ϕ which is usually given by $a(\phi) = \phi/\omega$ for a known constant ω, hence [4], $V(y) = b''(\theta)\phi/\omega$.

GLMs can be written in linear form of the parameters as,

$$\eta_i = g\big(E(y_i)\big) = g(\mu) = \boldsymbol{X_i'\beta}$$

and note that the expected response is given by,

$$E(y_i) = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{X_i'\beta})$$

If the link function, g, is chosen in such a way that $\eta_i = \theta_i$, then η_i is called canonical link [2]. Table 3.1 provides the canonical links ($\eta = \theta$) of some exponential family distributions. However, there are others that can be used as link function in GLM [2].

Table 3.1. *Canonical Links for exponential family distributions*

Distribution	$\boldsymbol{f(y)}$	$\boldsymbol{\theta}$	$\boldsymbol{\phi}$	$\boldsymbol{a(\phi)}$	$\boldsymbol{b(\theta)}$
Normal	$\frac{1}{\sigma\sqrt{2\pi}}exp(\frac{-(y-\mu)^2}{2\sigma^2})$	μ	σ^2	$\phi(=\sigma^2)$	$\frac{\theta^2}{2}$
Poisson	$\frac{\mu^y exp(-\mu)}{y!}$	$log(\mu)$	1	$\phi(=1)$	$exp(\theta)$
Binomial	$\binom{n}{y}(\frac{\mu}{n})^y(1-\frac{\mu}{n})^{n-y}$	$log(\frac{\mu}{1-\mu})$	1	$\phi(=1)$	$n\ log(1+e^\theta)$

Source: Wood, 2006 and Montgomery et al., 2012

Now, assuming the canonical link function is used, the likelihood of $\boldsymbol{\beta}$ (since y_i's are independent) is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i;\theta_i,\phi),$$

and hence, the log-likelihood function is given by,

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} log(f(y_i;\theta_i,\phi)) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + h(y_i,\phi) \end{aligned}$$

Thus, the equations to be solved in order to find estimate of β are given by

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} &= \frac{1}{\phi}\sum_{i=1}^{n} \frac{[y_i - b_i'(\theta_i)]}{b_i''(\theta_i)/\omega_i}\frac{\partial \mu_i}{\partial \beta_j} \\ &= \frac{1}{\phi}\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} \quad \text{for } \forall j \end{aligned}$$

Substituting Eq:3.2 yields,

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} = 0 \text{ for } \forall j \tag{3.3}$$

If $V(\mu_i)$ were known and independent from $\boldsymbol{\beta}$, the least square optimization objective would be

$$S = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{V(\mu_i)} \tag{3.4}$$

where, μ_i depends on $\boldsymbol{\beta}$ nonlinearly. Now, let $\hat{\boldsymbol{\beta}}^{[k]}$ be estimates at the k^{th} iteration, elements of vectors $\boldsymbol{\eta^{[k]}}$ and $\boldsymbol{\mu}^{[k]}$ are respectively given as $\eta_i^{[k]} = \boldsymbol{X_i\hat{\beta}^{[k]}}$ and $\mu_i^{[k]} = g^{-1}(\eta_i^{[k]})$. In each iteration [4], Eq:3.4 can be written as,

$$\begin{aligned} \boldsymbol{S} &= \left\| \sqrt{\boldsymbol{V_{[k]}^{-1}}}\boldsymbol{[y - \mu(\beta)]} \right\|^2 \\ &\approx \left\| \sqrt{\boldsymbol{W^{[k]}}}\boldsymbol{\left(z^{[k]} - X\beta\right)} \right\|^{\boldsymbol{2}} \end{aligned} \tag{3.5}$$

where, $V_{[k]ii} = V(\mu_i^{[k]})$, elements of the diagonal matrix $\boldsymbol{V}$. The pseudo-data, $z_i^{[k]}$ is,

$$z_i^{[k]} = g'(\mu^{[k]})(y_i - \mu_i^{[k]}) + \eta_i^{[k]}$$

and elements of the diagonal weight matrix $\boldsymbol{W^{[k]}}$ are given by

$$w_{ii}^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}$$

The method of iterative re-weighted least square (IRLS) [4] iterates to convergence where the converged $\hat{\boldsymbol{\beta}}$ solves Eq:3.3.

1. Using $\boldsymbol{\mu}^{[k]}$ and $\boldsymbol{\eta}^{[k]}$ obtain $\boldsymbol{z}^{[k]}$ and the weight matrix, $\boldsymbol{W}^{[k]}$.

2. Minimize Eq:3.5 with respect to $\boldsymbol{\beta}$ in order to obtain $\boldsymbol{\beta}^{[k+1]}$, and hence $\boldsymbol{\eta}^{[k+1]} = \boldsymbol{X\beta}^{[k+1]}$ and $\boldsymbol{\mu}^{[k+1]}$.

It can be noted that as initial values only $\boldsymbol{\mu}^{[0]}$ and $\boldsymbol{\eta}^{[0]}$ are needed but not $\hat{\boldsymbol{\beta}}^{[0]}$. Usually, initial values are taken as $\mu_i^{[0]} = y_i$ with slight adjustment in order to avoid infinite for $\eta_i^{[0]}$ (for example, when $y_i = 0$ with a log link) and $\eta_i^{[0]} = g(\mu_i^{[0]})$.

3.2 Generalized Additive Model as Penalized GLMs

In section 2.2.3, a basic additive model with two covariates was presented. As GLMs are to linear models, generalized additive models [7] are generalization of additive models.

The general structure of a GAM modeling a response variable, y with multiple predictors is,

$$g(\mu_i) = \boldsymbol{X_i^*\theta} + f_1(x_{1i}) + f_2(x_{2i}) + \dots \tag{3.6}$$

where, $\mu_i = E(y_i)$ and g is a function mapping y_i's through a mathematical transformation to the linear predictor, hence called a link function. In order the mapping to be ensured, a g must be twice differentiable monotonic function [4]. The most commonly used link functions are identity, logit, probit, and log [6]. In Eq:3.6, the GAM model has a parametric and smooth (nonparametric) components; $\boldsymbol{X_i}$ is i^{th} row of a model matrix for the parametric component with $\boldsymbol{\theta}$ being its vector of parameters, whereas, f_j is smoother of the predictor variable, x_j.

In literature, different approaches of GAM estimating method were proposed. Here, the smoothing spline basis approach which is a method incorporated in *mgcv::gam* [28] will be addressed.

One approach to estimate a GAM model is by choosing a basis for the smoothing function and a wiggliness measure. In this approach, model estimation implies estimation of smoothing parameter as well as model coefficients for a penalized likelihood maximization objective. This section puts forth the representation of a smoothing function using a basis and turn the GAM model into a penalized GLM, in which then the estimation will be accomplished in a similar way.

As discussed in section 2, given a basis function b_{ji} is chosen, a smoother, f_j, can be written as;

$$f_j(x_j) = \sum_{i=1}^{q_j} \beta_{ji} b_{ji}(x_j) \tag{3.7}$$

where, β_{ji}'s are parameters which need to be estimated. To construct the model matrix, let $\boldsymbol{\tilde{f}_j}$ be a vector where its j^{th} element is given by $\boldsymbol{\tilde{f}_j}_i = f_j(x_{ji})$ and vector of parameters of the j^{th} smoother be given by $\boldsymbol{\tilde{\beta}_j} = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jq_j}]$. Combining all together, it is easy to see that the model matrix for the j^{th} smoother is given by,

$$\boldsymbol{\tilde{f}_j} = \boldsymbol{\tilde{X}_j\tilde{\beta}_j} \tag{3.8}$$

where, $\boldsymbol{\tilde{X}_{j,ik}} = b_{jk}(x_{ji})$. As discussed in section 2.2.3, for bivariate additive model, Eq:3.6 suffers from identifiability problem. This can be avoided by constraining the sum or mean of $\tilde{\boldsymbol{f_j}}$ to zero [4].

$$1^T\tilde{\boldsymbol{X}}_{\boldsymbol{j}}\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = 0$$

By re-parameterization in concept of constraining [4], a matrix $\boldsymbol{Z}$ which satisfies the condition,

$$1^T\tilde{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{Z} = 0$$

and has the property that its $q_j - 1$ columns are orthogonal can be found. Writing $\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}} = \boldsymbol{Z}\boldsymbol{\beta}_{\boldsymbol{j}}$, by re-parameterizing the smooth in terms of $q_j - 1$ new parameters, $\boldsymbol{\beta}_{\boldsymbol{j}}$, a new model matrix for the j^{th} smoothing function given by $\boldsymbol{X}_{\boldsymbol{j}} = \tilde{\boldsymbol{X}}_{\boldsymbol{j}}\boldsymbol{Z}$, such that $\boldsymbol{f}_{\boldsymbol{j}} = \boldsymbol{X}_j\boldsymbol{\beta}_{\boldsymbol{j}}$ that satisfies the centering constraint will be obtained.

Once the model matrices of each smoothing function are centered (matrix), Eq:3.6 is written as

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i \tag{3.9}$$

where, $\boldsymbol{\beta}^T = [\boldsymbol{\theta}^{\boldsymbol{T}}, \boldsymbol{\beta}_{\boldsymbol{1}}^{\boldsymbol{T}}, \boldsymbol{\beta}_{\boldsymbol{2}}^{\boldsymbol{T}}, \boldsymbol{...}]$ and $\boldsymbol{X} = [\boldsymbol{X}^* \boldsymbol{:} \boldsymbol{X}_{\boldsymbol{1}} \boldsymbol{:} \boldsymbol{X}_{\boldsymbol{2}} : ...]$, a binded matrix of the parametric model matrix and all the centered ones.

3.2.1 Model parameter estimation

It can be seen now that Eq:3.9 is a GLM form, and its likelihood, say $l(\boldsymbol{\beta})$ can be written down in the same way as that of its GLM counterpart.

However, as it was discussed in section 2, if large q_j (number of knots) are used to represent the smoothers, f_j, and the method of maximum likelihood is used to estimate $\boldsymbol{\beta}$, the model parameters, then, there is possibility of over-fitting. This is the reason why penalized likelihood maximization is preferred over the ordinary likelihood maximization to estimate GAMs [4].

Given the roughness measure of each smoother, $\boldsymbol{\beta}\boldsymbol{S}_j\boldsymbol{\beta}$, the penalized likelihood is written as;

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{j}}\boldsymbol{\beta}, \tag{3.10}$$

Assuming the λ_j, smoothing parameters, are known, estimates of $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$, can be found by maximizing Eq:3.10 [4, 29].

For notational easiness, Eq:3.10 can be written as,

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{S}\boldsymbol{\beta}$$

where, $\boldsymbol{S} = \sum_j \lambda_j\boldsymbol{S}_j$. Now, by setting its derivatives with respect to $\boldsymbol{\beta}$ to zero, $l_p(\boldsymbol{\beta})$ can be maximized.

$$\frac{\partial l_p}{\partial \beta_j} = \frac{\partial l}{\partial \beta_j} - [\boldsymbol{S\beta}]_j = \frac{1}{a(\phi)}\sum_{i=1}^{n}\frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} - [\boldsymbol{S\beta}]_j = 0$$

In a similar way to that in section 3.1, the above system equation are those that would have to be solved to maximize the penalized non-linear least square optimization problem;

$$S_p = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{V(y_i)} + \boldsymbol{\beta}^T\boldsymbol{S\beta} \tag{3.11}$$

Again, assuming $V(y_i)$ are known, Eq:3.11 can be approximated as [4],

$$S_p \approx \left\|\sqrt{\boldsymbol{W}^{[k]}}\left(\boldsymbol{z}^{[k]} - \boldsymbol{X\beta}\right)\right\|^2 + \boldsymbol{\beta}^T\boldsymbol{S\beta} \tag{3.12}$$

where the vector of pseudo-data, $\boldsymbol{z}^{[k]}$ and values of the diagonal weight matrix $\boldsymbol{W}^{[k]}$ are respectively given by,

$$z_i^{[k]} = g(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + \boldsymbol{x_i}\hat{\boldsymbol{\beta}}^{[k]} \text{ and } w_{ii}^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}$$

Thus, given the smoothing parameters, λ_i's, the maximum penalized likelihood estimates, $\hat{\boldsymbol{\beta}}$, are obtained by repeating the following steps [4];

1. Given $\hat{\boldsymbol{\beta}}^{[k]}$, find $z^{[k]}$ and $w_{ii}^{[k]}$.
2. To find $\hat{\boldsymbol{\beta}}^{[k+1]}$, minimize Eq:3.12 with respect to $\boldsymbol{\beta}$. Repeat until convergence.

The same initial values are considered as that of IRLS method presented in section 3.1.

3.2.2 Degrees of freedom

Given the parameters are identifiable based on the data, degrees of freedom in ordinary linear regression model is equal to $tr(\boldsymbol{A})$ where $\boldsymbol{A}$ is the influence matrix, and that of the error term is $tr(\boldsymbol{I} - \boldsymbol{A}) = n - tr(\boldsymbol{A})$ [17].

In GAMs, the size of smoothing parameters involved in the process of penalized regression affects how many degrees of freedom a model will have. If the smoothing parameters (since there are several smoothing parameters in additive models) are all equal to zero then the fitted model would have degrees of freedom equal to the dimension of $\boldsymbol{\beta}$, on the other end, if the smoothing parameters are all too large, then the model will be over smoothed and this in turn results to a somewhat an inflexible model with few degrees of freedom [4].

Here, the so called effective degrees of freedom of a fit can be defined in various ways, one of which is using $tr(\mathbf{A})$ where $\mathbf{A}$ is the influence matrix. For the reason that different smoothing parameters would be used to smooth the penalty functions differently and affects the degrees of freedom differently, it is natural to look at the effective degrees of freedom by breaking down for each smooth. And this could be seen as equivalent to find the degrees of freedom for each model parameter, $\hat{\beta}_i$, since they are affected differently by the smoothing too.

From section 2.2.2, if we let $\boldsymbol{D} = \left(\boldsymbol{X^T X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X^T}$, then it follows that $\hat{\boldsymbol{\beta}} = \boldsymbol{Dy}$ and $\boldsymbol{A} = \boldsymbol{XD}$ which implies that $tr(\boldsymbol{A}) = tr(\boldsymbol{XD})$. Now, let $\boldsymbol{D}_i^0$ be equal to $\boldsymbol{D}$ when all it rows except the i^{th} row are zeroed. As a result, the elements of the vector $\boldsymbol{D}_i^0\mathbf{y}$ will be all zero except the i^{th} value which is $\hat{\beta}_i$. Consequently, trace of $\boldsymbol{A}$ can be written as;

$$tr(\boldsymbol{A}) = \sum_{i=1}^{D} tr(\boldsymbol{X}\boldsymbol{D}_i^0)$$

Therefore, $tr(\boldsymbol{X}\boldsymbol{D}_i^0)$ can be regarded as the effective degrees of freedom associated with the β_i. However, $tr(\boldsymbol{X}\boldsymbol{D}_i^0) = (\boldsymbol{DX})_{i,i}$. Now, if we define

$$\boldsymbol{R} = \boldsymbol{DX} = (\boldsymbol{X^T X} + \boldsymbol{S})^{-1}\boldsymbol{X^T X},$$

it can be seen that the leading diagonal of $\boldsymbol{R}$ is the vector of effective degrees of freedom of the model parameters and $tr(\boldsymbol{R}) = tr(\boldsymbol{A})$. Therefore, similar to that of the parametric models, the effective degrees of freedom of the model is given by $tr(\boldsymbol{A})$ and that of the residuals is $tr(\boldsymbol{I} - \boldsymbol{A})$ [4, 17].

To have an intuitive understanding about the effective degrees of freedom, it's worthy of relating it with that of an unpenalized estimates. It can be recalled from simple regression that the estimate of an unpenalized model is given by,

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$$

whereas, estimates of penalized model are given by

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X^T X} + \boldsymbol{S})^{-1}\boldsymbol{X^T y}\\ &= (\boldsymbol{X^T X} + \boldsymbol{S})^{-1}\boldsymbol{X^T X}(\boldsymbol{X^T X} + \boldsymbol{S})^{-1}\boldsymbol{X^T y}\\ &= \boldsymbol{R}\tilde{\boldsymbol{\beta}}\end{aligned}$$

Here, $\boldsymbol{R}$ is a mapping matrix between the unpenalized estimates and their corresponding penalized ones. Since, the unpenalized parameters have one degrees of freedom each, the term $\frac{\partial \hat{\beta}_i}{\partial \tilde{\beta}_i} = \boldsymbol{R}_{ii}$ is the change in the penalized parameter, $\hat{\beta}_i$, for a unit change of the unpenalized parameter, $\tilde{\beta}_i$. This means that the penalty involved in the model dwindles the degrees of freedom of the i^{th} term by a value of $\boldsymbol{R}_{ii}$, thus, $\boldsymbol{R}_{ii}$ is regarded as the effective degrees of freedom of the i^{th} model parameter.

3.2.3 Smoothing parameter selection

The problem of choosing an appropriate smoothing parameter is omnipresent in fitting a curve. Choosing the degree, in polynomial regression estimation and setting the basis dimension in regression spline are equivalent to the choice of a smoothing parameter. Figure 3.1 shows the effect of smoothing parameter in the penalized regression spline estimation technique at different values. If λ is too small, the fit will be too wiggly and if λ is too large, it will be over smoothed where in both cases the estimated spline $\hat{f}$ can't approximate the true function f. Instead of arbitrarily picking value of λ, it would be good to use some techniques to find the optimum value so that $\hat{f}$ will be as close as possible to f yet a better one though there are other suggested estimators.

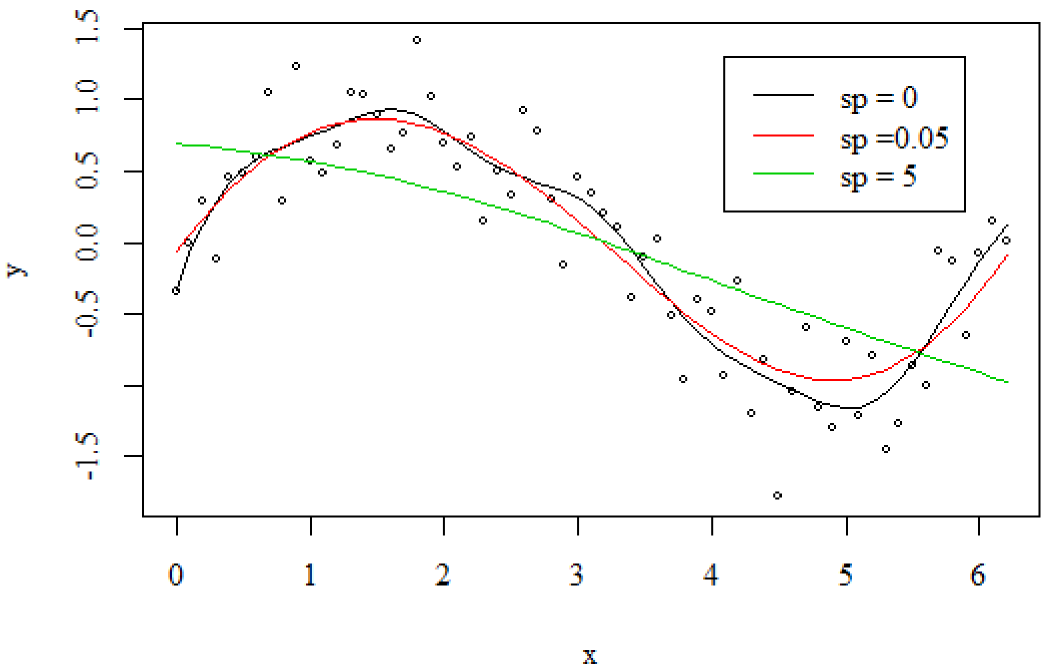


Figure 3.1. *Effect of smoothing parameter in model fitting using penalized regression spline (as the value of the smoothing parameter increases, the fit approaches to a straight line)*

There are two different and somehow opposing philosophical approaches into the smoothing parameter choosing problem [17]. The first one is a subjective choice of the parameter which hypothetically allows to explore the features of the data, and the second approach is that there should be an automatic method by which the data chooses the smoothing parameter by itself. Cross-validation is one of the most commonly used automatic ways of selecting a smoothing parameter.

3.2.3.1 *Ordinary cross validation*

In curve estimation, taking into consideration that the error term has zero mean, the best predictor of y_i is $f(x_i)$, where $f(x)$ is the ideal curve. Consequently, the best estimator $(\hat{f}(x)$ of the true curve is the one that minimizes the term $(y - \hat{f}(x))^2$ for a ”new value” y at a given point x. This is the fundamental motivation for using cross-validation to choose smoothing parameter [17]. Theoretically, acceptable criteria to obtain λ so that $\hat{f}$ minimizes

$$M = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i - f_i\right)^2. \tag{3.13}$$

In reality, there is no ”new value/observation” when a smoothing is applied to the whole dataset. In cross validation, however, the smoothing curve is estimated by leaving one observation (y_i, x_i) out in order to have that ”new observation” used for prediction and this one-leave-out estimator is denoted by $\hat{f}^{-1}$. As in every other regression models, how good $\hat{f}^{-1}$ in predicting a new observation is possibly determined by how close it is to y_i.

Now, let $\hat{f}^{-1}$ be the model fitted to the remaining data when y_i is left out. Ordinary cross validation (OCV) is then given by,

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{-1} - y_i\right)^2 \tag{3.14}$$

Replacing y_i by $f_i + \epsilon_i$,

$$\begin{aligned}\nu_o =& \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{-1} - f_i - \epsilon_i\right)^2\\ =& \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{-1} - f_i\right)^2 - 2\left(\hat{f}_i^{-1} - f_i\right)\epsilon_i + \epsilon_i^2.\end{aligned}$$

For the fact that $E(\epsilon_i) = 0$ and ϵ_i and $\hat{f}_i^{-1}$ are independent, the expected value of the second term is zero. Consequently,

$$E(\nu_o) = \frac{1}{n} E\left(\sum_{i=1}^{n} \left(\hat{f}_i^{-1} - f_i \right)^2 + \epsilon_i^2 \right)$$

With the concept of large sample, it's fairly true that $\hat{f}^{-1} \approx f$ which in turn results to the conclusion that $E(\nu_o) \approx E(M) + \sigma^2$. Thus, it is a reasonable approach to choose λ that minimizes ν_o.

Calculating ν_o appears to be tedious for the reason that n separate smoothing curves ($\hat{f}^{-1}$) have to be fitted in order find the OCV score value. However, there is a simplified way of obtaining ν_o using the influence or hat matrix of the penalized regression model [17, 30] which is given by,

$$\nu_o = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}_i}{1 - A_{ii}} \right)^2 \tag{3.15}$$

where, $\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X^T X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X^T}$ is the influence matrix and $\hat{f}$ is the smoothing spline obtained from all the data. Detailed proof of Eq:3.15 can be found in Green and Silverman [17] and Wahba [30].

In Eq:3.14, given the diagonal values, A_{ii}, are provided, OCV score can easily be calculated from the residuals, $(y - \hat{f})$, of the spline obtained by smoothing all the data.

3.2.3.2 *Generalized cross validation*

The problem with ordinary cross validation is that, firstly it is computationally expensive especially in the additive case where there are more than one smoothing

parameters to be estimated and another is, given the penalized regression optimization problem for additive models which is given by

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \sum_{i=1}^{p} \lambda_i \boldsymbol{\beta^T S_i \beta},$$

it normally should have identical solutions in terms of $\boldsymbol{\beta}$ even when for any orthogonal matrix $\boldsymbol{Q}$, it is rotated as,

$$\|\boldsymbol{Qy} - \boldsymbol{QX\beta}\|^2 + \sum_{i=1}^{p} \lambda_i \boldsymbol{\beta^T S_i \beta}.$$

However, these two optimization problems generally results in different OCV scores, which is labeled as invariance problem of OCV [30].

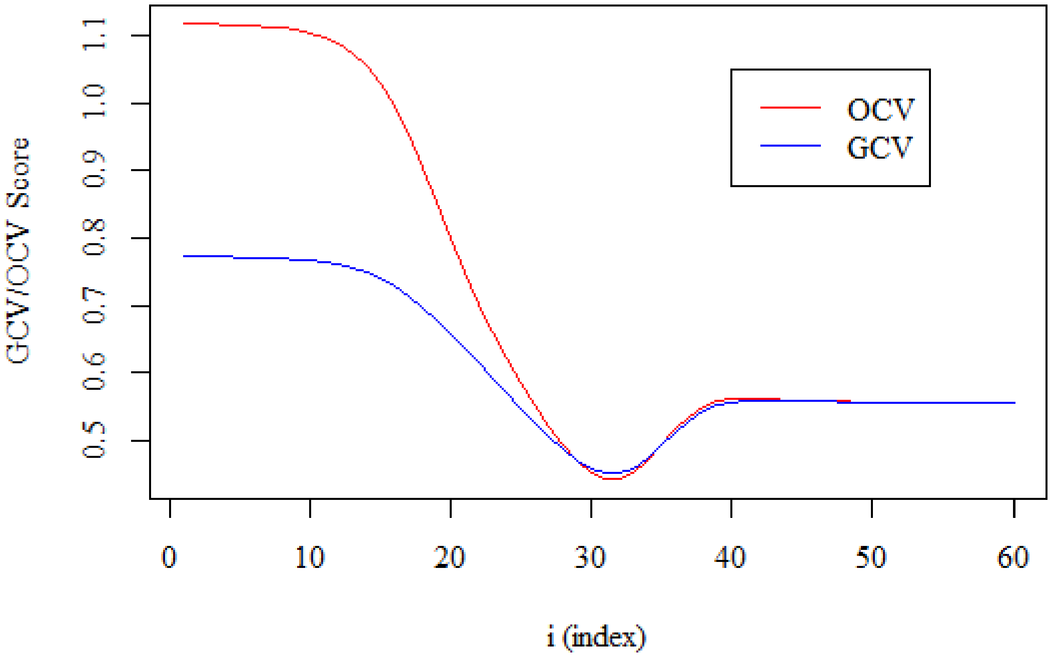


Figure 3.2. *Optimal smoothing parameter using ordinary cross validation and generalized cross validation;* $\lambda = 1.5^i * 10^{-8}$
Source: *Wood, 2006; p. 131*

The approach of generalized cross validation (GCV) is an extension of the ordinary cross validation technique in which the weights $(1 - A_{ii})$ are replaced by their average value, $tr(\boldsymbol{I} - \boldsymbol{A})/n$. The GCV score is then obtained as,

$$\nu_g = n \frac{\sum_{i=1}^{n} \left(y_i - \hat{f} \right)^2}{[tr(\boldsymbol{I} - \boldsymbol{A})]^2} \tag{3.16}$$

Figure 3.2 shows that ν_o and ν_g were optimized at different iterations. And, the smoothing parameters, λ, from the two methods were used to fit smoothing splines shown in Figure 3.3 in which a slightly differing models were obtained for the univariate dataset used.

3.2.4 Model deviance

Model evaluation plays a fundamental role in regression analysis; comparisons can be made between models to obtain the better of them. In linear regression, mean square error (MSE) is regarded as the building blocks of most model evaluation techniques

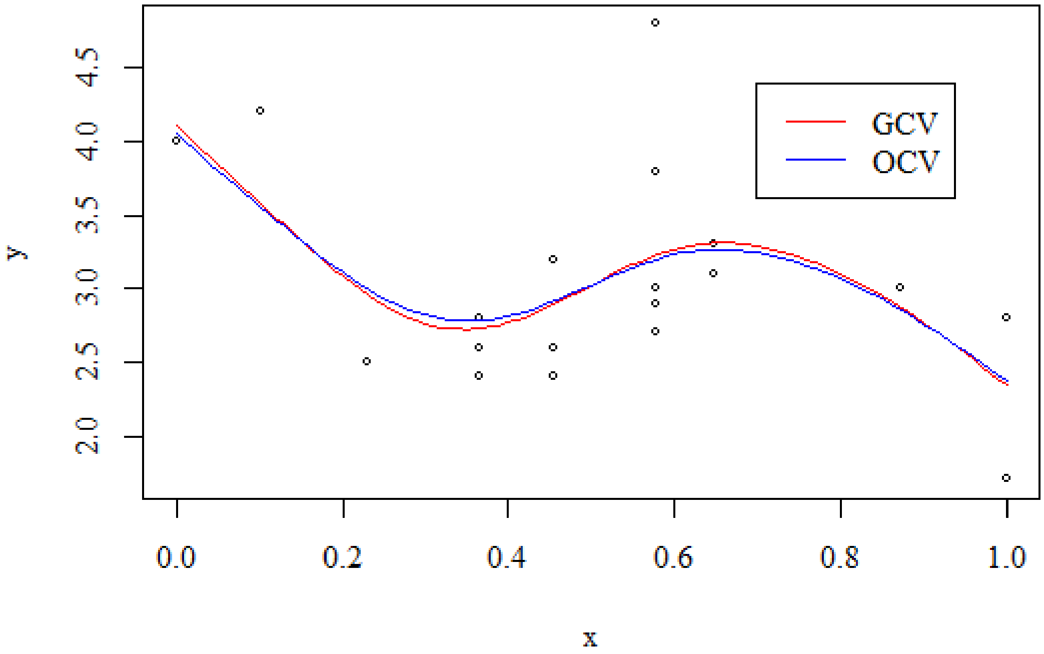


Figure 3.3. *GCV and OCV optimal fit*

and inferences made for it measures how far the model estimations from the actual observations are. In GLMs and GAMs, it is necessary to have a quantity which is equivalent in importance and interpretation to residual sum of squares for ordinary linear modeling [4].

As minimizing MSE is to least square fits, in models fitted using maximum likelihood estimation (MLE), the quantity to be minimized is the deviance. Maximizing the likelihood in those models corresponds to minimizing the deviance of the model [27]. Model deviance is defined as twice the difference in log-likelihood between the saturated model and the full model (model of interest) [2, 31, 32]. It is given by,

$$\begin{aligned} D =& 2[l(\hat{\boldsymbol{\beta}}_{max}) - l(\hat{\boldsymbol{\beta}})]\phi \\ =& \sum_{i=1}^{n} 2\omega_i\big[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\big] \end{aligned} \tag{3.17}$$

where $l(\hat{\boldsymbol{\beta}})_{max}$ is maximum likelihood of the saturated model: the model which have separate parameter for each observation and a perfect fit, $\hat{\boldsymbol{\mu}} = \boldsymbol{y}$ [4, 27]. $\tilde{\theta}$ and $\hat{\theta}$ are respectively the maximum likelihood estimates of canonical parameters which are provided in Table 3.1 for the saturated model and model of interest [4]. ω is a constant which in most cases is equal to 1. Deviance of a model can be regarded as the lack of fit between the model and the data points. It is used for model adequacy checking; the smaller the deviance is the better the model.

3.3 Multicollinearity

In multiple regression analysis, if the predictor variables are orthogonal, then inference can relatively be easily done. By orthogonality, it means that the regressors have no linear relationship. However, multicollinearity, the presence of near-linear dependencies among the predictors, has serious effects on the least square estimates of the model. Montgomery et al., 2012 puts forth the effects of multicollinearity; it overestimates the variance of the estimators, $V(\hat{\beta})$ as well as the absolute values of the estimators.

Consider a linear model of y with two regressors, x_1 and x_2 which is given by;

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{3.18}$$

and let all the variables are scaled to a unit length [2]. The normal equations of the least-squares approach are

$$(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$
$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is correlation coefficient between x_1 and x_2. Similarly, the correlation between y and the predictors are respectively given by r_{1y} and r_{2y}. The inverse of the matrix $\boldsymbol{X}'\boldsymbol{X}$ is

$$\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}.$$

Hence, the estimates of the model coefficient are given by

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2} \text{ and } \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2} \tag{3.19}$$

In matrix form, the variance of model coefficients is $V(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{C}\sigma^2$. Thus, the individual variance of $\hat{\beta}_1$ and $\hat{\beta}_2$ are respectively given by $V(\hat{\beta}_1) = C_{11}\sigma^2$ and $V(\hat{\beta}_2) = C_{22}\sigma^2$. In general, the variance of $\hat{\beta}_i$ in a linear regression model with multiple predictors is written as $V(\hat{\beta}_i) = C_{ii}\sigma^2$ and covariance of $\hat{\beta}_i$ and $\hat{\beta}_j$ is $Cov(\hat{\beta}_i, \hat{\beta}_j) = C_{ij}\sigma^2$.

In the existence of multicollinearity, the correlation between x_1 and x_2, r_{12}, will be large which in turn affects the variance of model coefficients. if $|r_{12}| \to 1$ then $V(\hat{\beta}_1) \to \infty$, $V(\hat{\beta}_2) \to \infty$ and $|Cov(\hat{\beta}_i, \hat{\beta}_j)| \to \infty$. This makes clear that existence of multicollinearity produces model coefficients with inflated variance. Furthermore, Montgomery et al., 2012 put forth that the least-square estimates, $\hat{\beta}_j$, are too large in absolute value if multicollinearity exists.

3.4 Outlier

Outlier is an abnormal observation which differ greatly from the rest of the data. In parametric regression analysis, presence of outliers in dataset disturbs the quality of least-square estimates [2] because the optimization is to minimize the squared deviations. Least-square estimates can be influenced by abnormal observations both in the response variable as well as in the predictor variables. Data points which are remote from the rest of the data in terms of some values of the regressors while the value of the response variable is consistent are referred as leverage points. In this paper, only outliers (extreme points in the response variable) and the consequence they have in regression analysis will be considered.

4. SIMULATION

This section discusses the simulation studies performed to explore the performance of GAM models based on three different smoothing spline bases under multicollinearity and outliers. Specifically, data were generated with and without outliers with different predictor functions and proportion and location of outliers and different degrees of multicollinearity.

4.1 Data with Outliers

In this section, a simulation study is provided to compare the performance of GAM models for binomial and Poisson response variables. Recall that for data $\{(x_j, y_j), j = 1, 2, ..., n\}$, univariate additive model is given by

$$y_j = f(x_j) + \epsilon_j \tag{4.1}$$

where f is one of cubic spline, p-spline or thin-plate spline.

Four scenarios were considered to generate the data $\{(x_j, y_j), j = 1, 2, ..., n\}$ using the functions adopted from [11] and [12].

Scenario 1: The covariate X follows a uniform distribution $X \sim U(0, 1)$. The response variable, Y, was simulated from the distribution, $Y/X \sim Poisson(\lambda(X))$, where $\lambda(X) = g^{-1}(h_1(X))$. Here, g is a log-link function.

$$h_1(X) = 4cos(2\pi(1 - X)^2) \tag{4.2}$$

For a specified outlier proportion value given by $\delta : 0, 0.1, 0.2$, a total of $n\delta$ were randomly selected to be changed to outliers in the following manner. The randomly selected Y-values were multiplied by $u_1^{u_2}$ as given in Eq:4.3, and the result was rounded to the nearest integer.

$$Y = Yu_1^{u_2} \tag{4.3}$$

where, $u_1 \sim U(2, 5)$ and $u_2 \in (-1, 1)$.

Scenario 2: Here, a binomial response variable was generated from the distribution $Y/X \sim Binomial(1, p(X))$ with $p(X) = g^{-1}(h_1(X))$ where g is the logit-link function.

Covariate X is generated as given in scenario 1. The proportion of outliers given in scenario 1 are also applied here without any change. Outliers were included to the response variable in such a way that if the randomly selected value of Y is 1, then it is replaced by 0 otherwise by 1.

Scenario 3: In this scenario, the covariate is given by $X = i$ for $i = 1, 2, ..., n$ and response variable is generated from the distribution
$Y/X \sim Poisson(\lambda(X))$ where $\lambda(X) = g^{-1}(h_2(X))$ and g is a log-link function.

$$h_2(X) = sin(2X/120) + cos(7X/60) + 1 \tag{4.4}$$

In this case, outliers was included into the response variable by using Eq:4.5.

$$y_j = (1 - z_j)y_j + z_j w_j, \quad j = 1, 2, ..., n \tag{4.5}$$

where $z_j \sim Binomial(1, \delta)$ and $w_j \sim Poisson(30)$.

Scenario 4: The covariate X used in scenario 3 was also used here as it is. However, here a binomial response variable was simulated from the distribution $Y/X \sim Binomial(10, p(X))$. The parameter p is given by $p(X) = g^{-1}(h_3(X))$ where g is the logit link function.

$$h_3(X) = -sin(5X/120)/0.8 - 1 \tag{4.6}$$

For including outliers in the response variable, the procedure given by Eq:4.5 was used. All the settings used in scenario 3 for outlier inclusion were kept fixed except $w_j = 10$.

Table 4.1. *Generating parameter for Poisson and binomial response variables*

Distribution	Parameter	
Poisson	$\lambda = exp(h_i(x))$	$i = 1, 2$
Binomial	$p = \frac{exp(h_i(x))}{1+exp(h_i(x))}$	$i = 1, 3$

In scenarios 3 and 4, the number of outliers included in the response variables is random which is controlled by the values of δ [11].

4.2 Data with Multicollinearity

This section presents a simulation study for generating data with varying degrees of multicollinearity between covariate variables. To generate data with a desired degrees of multicollinearity, the simulation design given by McDonald and Galarneau, 1975 is adopted.

First, independent standard normal pseudo random numbers, z_{ij} for $i = 1, 2, ..., n$ and $j = 1, 2, ..., 5$ were generated. Then, Eq:4.7 was used to generate a total of four covariates with a specified degree of linear relationship.

$$x_{ij} = (1 - \rho^2)^{1/2} z_{ij} + \rho z_{i5}, \quad i = 1, 2, ..., n; \; j = 1, 2, 3, 4. \tag{4.7}$$

where ρ is specified so that the correlation between any two covariates will be approximately equal to ρ^2. Here, three different values of ρ (0.9, 0.99, 0.999) were considered.

The response variable is then generated using

$$Y = 2 + 5X_1 + 3X_2 + 4X_3 + 8X + \epsilon \tag{4.8}$$

where ϵ is the error term which is generated from $N(0, \sigma^2)$. In this study, four values of the error variance ($\sigma^2 = 1, 9$, 25 and 100) were used. The coefficients given in Eq:4.8 are arbitrary.

In order to understand how these techniques behave with varying sample size, data with size of $n = 50, 100, 500$ were simulated under the above specified constraints of degrees of linear relationship and error variances. Here, it is important to note that the reason for not using a smaller sample size ($n < 50$) is that the number of parameters to be estimated are exceeding the sample size.

4.3 Data with Both Outliers and Multicollinearity

This section addresses a simulation study for generating data with linearly related covariates and a response variable containing outliers. All the settings used in section 4.2 are carried over to this section except that outliers are included to the response variable, Y. The inclusion of outliers was achieved by randomly selecting 10% of response values and multiplying them by 20 in order to inflate them in their absolute values. For the same reason discussed in section 4.2, the sample sizes used here are $n = 50, 100, 500$.

5. RESULTS AND DISCUSSION

This section presents the results of the analysis conducted in the study. Data generated using the procedures discussed in section 4 were used to evaluate the performance of the three smoothing splines; cubic regression, p-spline and thin plate spline. For the analysis and model building procedures, the $mgcv$ package [28] in R statistical software was used.

It is important to note that, in each scenario mentioned in section 4.1, sample sizes of $20, 50, 100, 200$, and 500 were simulated and for each sample size a total of 500 repetitions were generated. Each of the three smoothing splines were used to fit a GAM model for all the samples. Then after, the mean and standard deviation of the model deviances for each smoother were obtained. In addition, the proportion of the number of times a model resulting in the smallest deviance was obtained.

5.1 Performance of Models in the Presence of Outlier

Mean and standard deviation of deviances of the models fitted using cubic regression spline(cr), p-spline (ps) and thin-plate spline (tp) bases are presented in Table 5.1 and Table 5.2. Additionally, proportion of the number of times each model produced smaller deviance is provided. In cases where response variable is Poisson, data generated using scenario 1 and scenario 3 of section 4 are used, whereas, for the binomial response case data are simulated using the procedures given in scenarios 3 and 4.

For the cases where Poisson response variable is considered, the results of the experiments are provided in Table 5.1. The results show that an increase in the number of outliers included in the response variable has inflated the mean deviances of all the models. To illustrate this, one sample size (e.g n = 20) can be considered for comparison of the outcomes when different number of outliers are included in the data. In cases where an outlier is not included, the mean deviances of cubic, p and thin plate splines are respectively $9.09, 9.17$, and 8.91, however, when some outliers (δ = 0.1) are introduced in the response variable their respective mean deviances are $37.62, 36.3$, and 35.37 respectively. Similarly, for δ = 0.2 as well, the mean deviances are seen to be increasing. Similarly, results obtained using scenario 3 supports this argument.

Results from scenario 1 demonstrate that for $n = 20$ thin plate spline produced a smaller mean deviance regardless of the outlier proportion included. For larger sample sizes ($n \neq 20$) however, p-spline performed better in all combinations of outlier

Table 5.1. *Mean, standard deviation and proportion of the number of times a model resulted in a small deviance for a Poisson response variable with outliers*

		cr		**ps**		**tp**		**cr**	**ps**	**tp**
δ	n	MD	SD	MD	SD	MD	SD		P	
				Scenario 1						
0	20	9.09	4.35	9.17	4.32	8.91	4.34	0.33	0.29	0.38
	50	33.21	9.41	33.01	9.05	33.08	9.47	0.32	0.39	0.29
	100	74.63	13.38	72.95	12.51	74.43	13.27	0.21	0.59	0.2
	200	159.9	19.46	155.48	18.38	159.75	19.33	0.11	0.79	0.09
	500	414.72	29.71	402.13	28.05	414.58	29.64	0.03	0.93	0.04
0.1	20	37.62	44.29	36.3	45.53	35.37	43.51	0.34	0.36	0.3
	50	187.22	126.95	181.54	124.81	183.95	125.08	0.27	0.44	0.29
	100	456.72	224.2	443.83	215.01	451.23	219.91	0.17	0.53	0.3
	200	971.63	327.34	956.48	323.52	967.13	325.73	0.14	0.65	0.21
	500	2609.78	552.51	2583.97	548.34	2605.96	551.26	0.09	0.76	0.15
0.2	20	64.37	64.85	61.78	70.41	59.38	63.98	0.3	0.36	0.33
	50	325.57	173.34	316.41	175.07	320.13	174.73	0.21	0.44	0.35
	100	797.04	284.76	777.25	279.71	788.55	284.35	0.15	0.58	0.27
	200	1767.44	437.71	1741.02	434.11	1758.91	437.08	0.13	0.64	0.23
	500	4690.62	705.33	4653.87	698.59	4683.45	704.91	0.13	0.71	0.16
				Scenario 3						
0	20	16.35	5.79	16.66	5.85	17.14	6.19	0.31	0.39	0.3
	50	48.04	9.65	48.21	9.57	48.04	9.64	0.32	0.37	0.31
	100	98.12	14.7	98.39	14.58	98.15	14.73	0.37	0.34	0.28
	200	243.33	24.6	343.27	32.8	209.8	21.5	0	0	1
	500	1565.1	70.98	1559.9	70.69	1564.21	70.88	0.05	0.72	0.23
0.1	20	46.72	28.19	48.07	29.32	46.44	27.49	0.41	0.35	0.24
	50	240.68	85.29	249.89	89.56	241.49	85.76	0.45	0.27	0.28
	100	494.37	115.59	497.91	116.38	494.7	115.56	0.41	0.37	0.23
	200	1123.61	185.11	1194.98	188.11	1100.93	186.17	0.13	0.04	0.83
	500	3992.12	316.7	3994.41	316.41	3992.6	316.65	0.33	0.4	0.27
0.2	20	77.48	38.15	79.11	37.16	77.18	37.83	0.4	0.37	0.23
	50	378.56	91.91	391.63	91.87	380.52	92.15	0.48	0.25	0.28
	100	756.26	113.98	761.37	112.19	756.98	113.46	0.4	0.36	0.24
	200	1670.21	164.73	1717.14	163.7	1655.21	164.64	0.21	0.08	0.72
	500	5419.24	271.15	5420.84	269.46	5419.02	271.62	0.3	0.42	0.29

MD: Mean deviance; SD: Standard Deviation

P: proportion of the number of times a model produced smaller deviance.

proportion (δ) and sample sizes. In addition to this, the results obtained from p-spline where more consistent; In most cases standard deviation of model deviances produced using p-spline are found to be smaller. Furthermore, the proportion of smaller deviance scored by each method is another evidence that p-spline outperformed the others. The proportion of p-spline is higher in all the cases except for the combination of $\delta = 0$ and $n = 20$. Moreover, it can be noted that dominance of p-spline increases with the increase of sample size. For example, for $\delta = 0.1$, p-spline produced smaller deviance only 36% of the times, however, for $n = 500$ and $\delta = 0.1$, 76% of the time it produced a smaller deviance.

On the contrary, results from scenario 3 do not show the same trend as that of scenario 1, rather, p-spline seems to be outperformed by both cubic and thin plate splines. Almost in all the cases, the mean deviance obtained from it was found to be larger than that of the others. Despite this, it produced the smallest standard deviation in majority of the cases which indicates that the models fitted using p-spline for the 500 samples were more consistent among each other than those fitted using either cubic regression spline or thin plate spline.

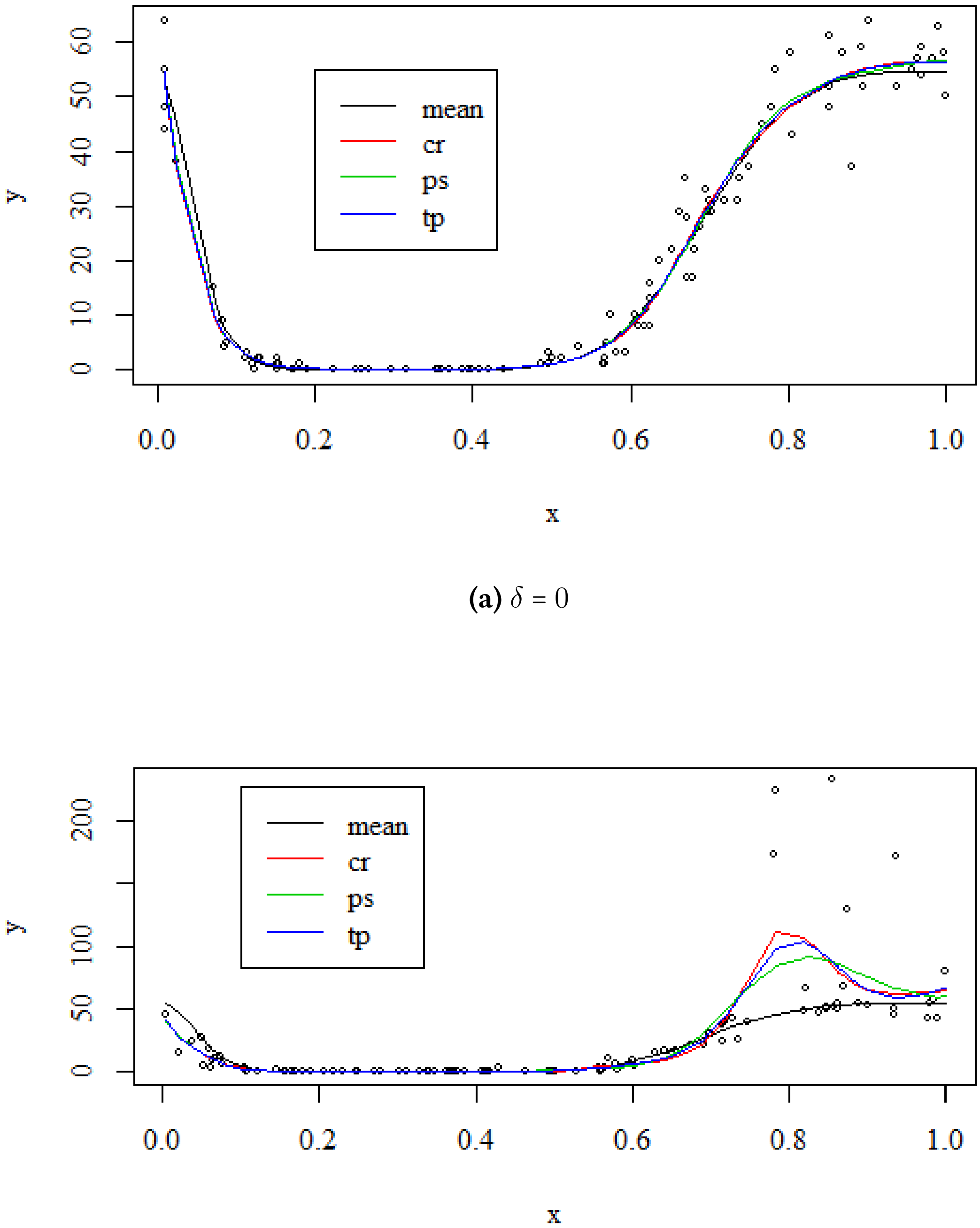
Looking at the proportions of scenario 3, the number of small deviances produced by each method is almost the same when there is no outlier and $n \leq 100$, where as, in the presence of outliers, proportion of ”smaller deviance” of cubic regression is larger in most cases. Nevertheless, when a sample size of $n = 200$ is considered, thin plate spline dominates the others in an exceptional way; it produced smaller deviance 100% when $\delta = 0$, 83% when $\delta = 0.1$ and 72% of the times when $\delta = 0.2$.

Figure 5.1 and Figure 5.2 display the models fitted for Poisson data simulated with an without outliers using scenarios 1 and 3 respectively for $n = 100$. In both cases, the fit of the three models have similar trend. It is however important to note that the presence of outliers have an impact on the model fit. All the models tend to depart from the true mean in positions where outlier is present.

Table 5.2 shows the mean deviances of the three models of interest along with the proportions of their small deviance when a binomial response variable is considered. The first block under Scenario 2 provides the outputs when the models are applied to fit data generated using Scenario 2 of section 4 while, the second block is for the results obtained when the data simulated using Scenario 4 is used.

As in the case of Poisson response variable, presence of abnormal observations in binomial response inflates the deviance of the models. For elaboration, take the case for $n = 20$ under scenario 2 of Table 5.2. When data free of outliers are used, the obtained mean deviances in order of their occurrence in the table are $2.53, 3.69$, and $3, 45$, whereas for $n = 20$ and $\delta = 0.2$, model deviances are 8.88, 13.23 and 14.38 respectively. Similarly, in the case of the second simulation, for the same combinations of n and δ, the mean deviation increased from $15.67, 15.78$, and 16.12 at $\delta = 0$ to $32, 34.25$ and 34.16 for $\delta = 0.2$. It is important however not to compare the magnitude of the change in cases of Poisson and binomial response variables for they have different units.

In situations where no outlier is included for data simulated using scenario 2, the

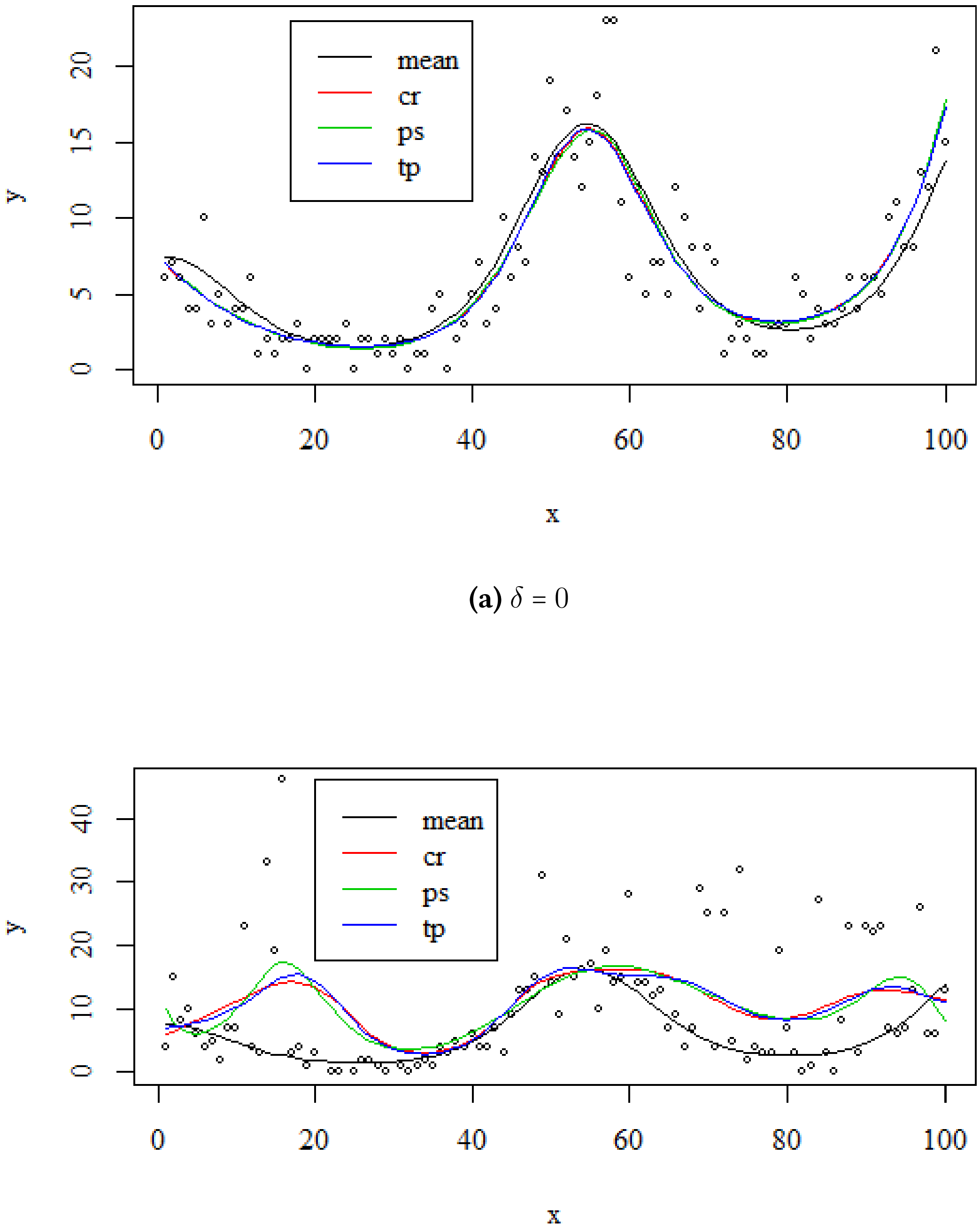


(a) δ = 0

(b) δ = 0.2

Figure 5.1. *GAM models fitted using cubic regression, p-spline and thin-plate spline for data with Poisson response which is generated using Scenario 1*

proportion of the models are close to each other regardless of the size of n. With the increase of outlier proportion and sample size however, the frequency of p-spline looks inferior to that of the rest spline techniques under consideration. Cubic regression and



(a) $\delta = 0$

(b) $\delta = 0.2$

Figure 5.2. *GAM models fitted using cubic regression, p-spline and thin-plate spline for data with Poisson response which is generated using Scenario 3*

thin plate splines produced relatively similar deviances in most cases. If the case of scenario 4 is considered, remarkably, cubic regression outperformed both p-spline and thin plate spline in the sense of producing smaller mean deviance. Generally speaking, despite producing better results in the case of Poisson response, p-spline was found to be the least best when a binomial response is used.

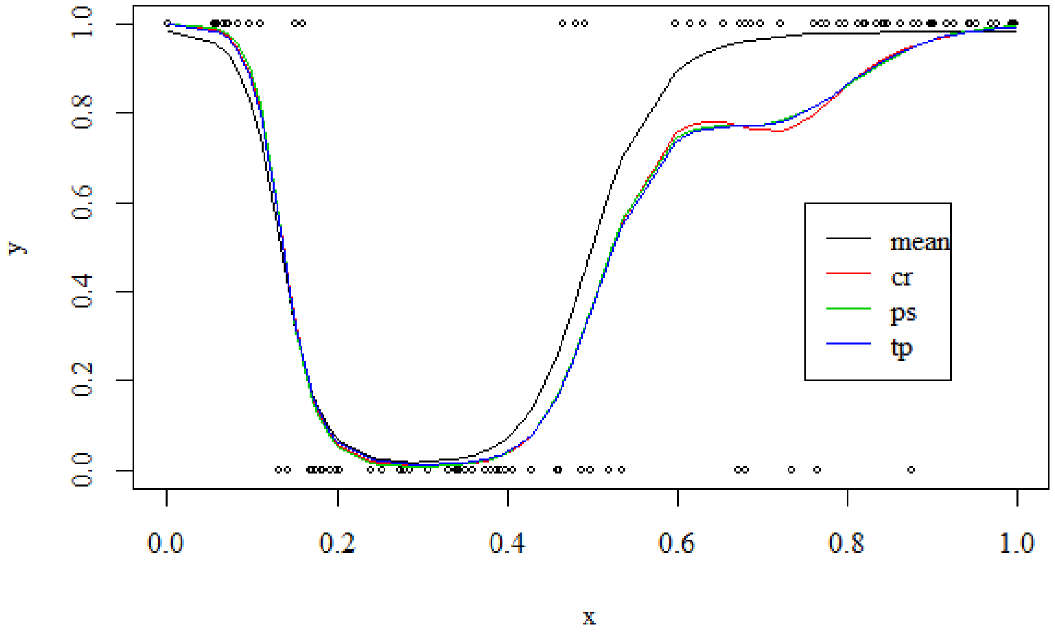
Table 5.2. *Mean, standard deviation and proportion of the number of times a model resulted in a small deviance for a binomial response variable with outliers*

		cr		ps		tp		cr	ps	tp
δ	n	MD	SD	MD	SD	MD	SD		P	
					Scenario 2					
0	20	2.53	3.85	3.69	4.66	3.45	4.52	0.37	0.34	0.29
	50	16.63	10.31	17.65	10.18	16.15	10.56	0.33	0.35	0.32
	100	42.98	13.89	43.56	13.53	42.67	14.02	0.34	0.33	0.34
	200	94.84	17.7	95.73	17.64	94.99	17.58	0.33	0.35	0.32
	500	247.74	28.56	248.28	28.52	247.71	28.43	0.35	0.35	0.3
0.1	20	5.38	6.54	8.54	7.1	8.6	7.44 0	0.44	0.34	0.22
	50	34.22	10.32	36.03	9.47	34.73	10.62	0.39	0.33	0.28
	100	81.61	9.51	82.1	9.46	81.55	9.76	0.33	0.32	0.35
	200	170.89	13.02	171.82	12.79	170.81	12.96	0.37	0.26	0.37
	500	435.27	18.54	436.46	18.76	435.16	18.74	0.37	0.21	0.42
0.2	20	8.88	8.4	13.23	8.25	14.38	8.39	0.48	0.37	0.15
	50	48.21	8.5	49.31	8.09	49.09	8.4	0.31	0.41	0.28
	100	105.94	7.62	106.44	7.47	106.07	7.7	0.3	0.33	0.37
	200	218.04	9.45	218.73	9.49	217.95	9.48	0.35	0.25	0.39
	500	554.97	14.45	556.41	14.13	554.91	14.4	0.4	0.19	0.42
					Scenario 4					
0	20	15.67	4.75	15.78	4.76	16.12	4.87	0.32	0.41	0.27
	50	47.2	7.54	47.39	7.57	47.82	7.69	0.26	0.47	0.26
	100	87.77	11.37	87.98	11.45	87.88	11.42	0.27	0.42	0.31
	200	164.9	15.7	165.21	15.71	164.86	15.69	0.28	0.44	0.28
	500	414.93	24.57	481.06	28.29	411.01	24.67	0.19	0	0.81
0.1	20	23.97	9.37	25.81	10.74	25.75	10.37	0.4	0.34	0.26
	50	108.31	26.7	110.91	27.91	110.52	27.38	0.45	0.3	0.25
	100	202.99	35.34	206.29	35.78	204.47	35.65	0.46	0.25	0.29
	200	387.06	47.63	387.18	47.63	387	47.56	0.34	0.38	0.28
	500	942.57	68.78	984.95	68.93	939.87	68.74	0.31	0.02	0.67
0.2	20	32	12.28	34.25	13.97	34.16	14.04	0.4	0.34	0.26
	50	150.18	27.93	153.36	29.47	152.04	29.24	0.4	0.34	0.27
	100	279.8	35.9	283.17	35.75	281.12	36.11	0.44	0.28	0.28
	200	526.96	41.88	526.15	42.27	526.79	41.82	0.31	0.43	0.25
	500	5419.24	271.15	5420.84	269.46	5419.02	271.62	0.3	0.42	0.29

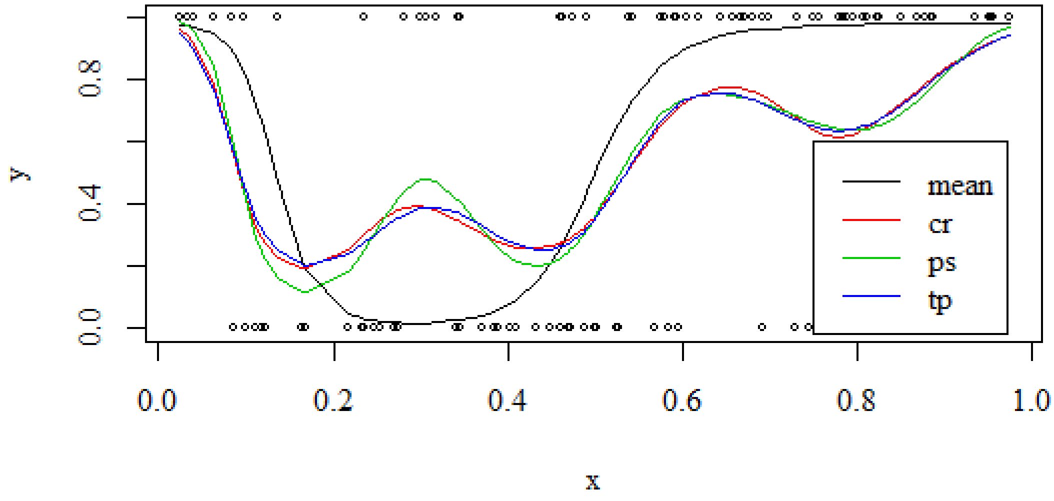
MD: Mean deviance; SD: Standard Deviation

P: proportion of the number of times a model produced smaller deviance

Figure 5.3 and Figure 5.4 show GAM models fitted using cubic regression (cr), p-spline (ps) and thin-plate spline (ps) for data simulated using Scenarios 2 and 4 respectively. For the sake of illustration only the case when $n = 100$ and $\delta = 0.2$ is shown here.

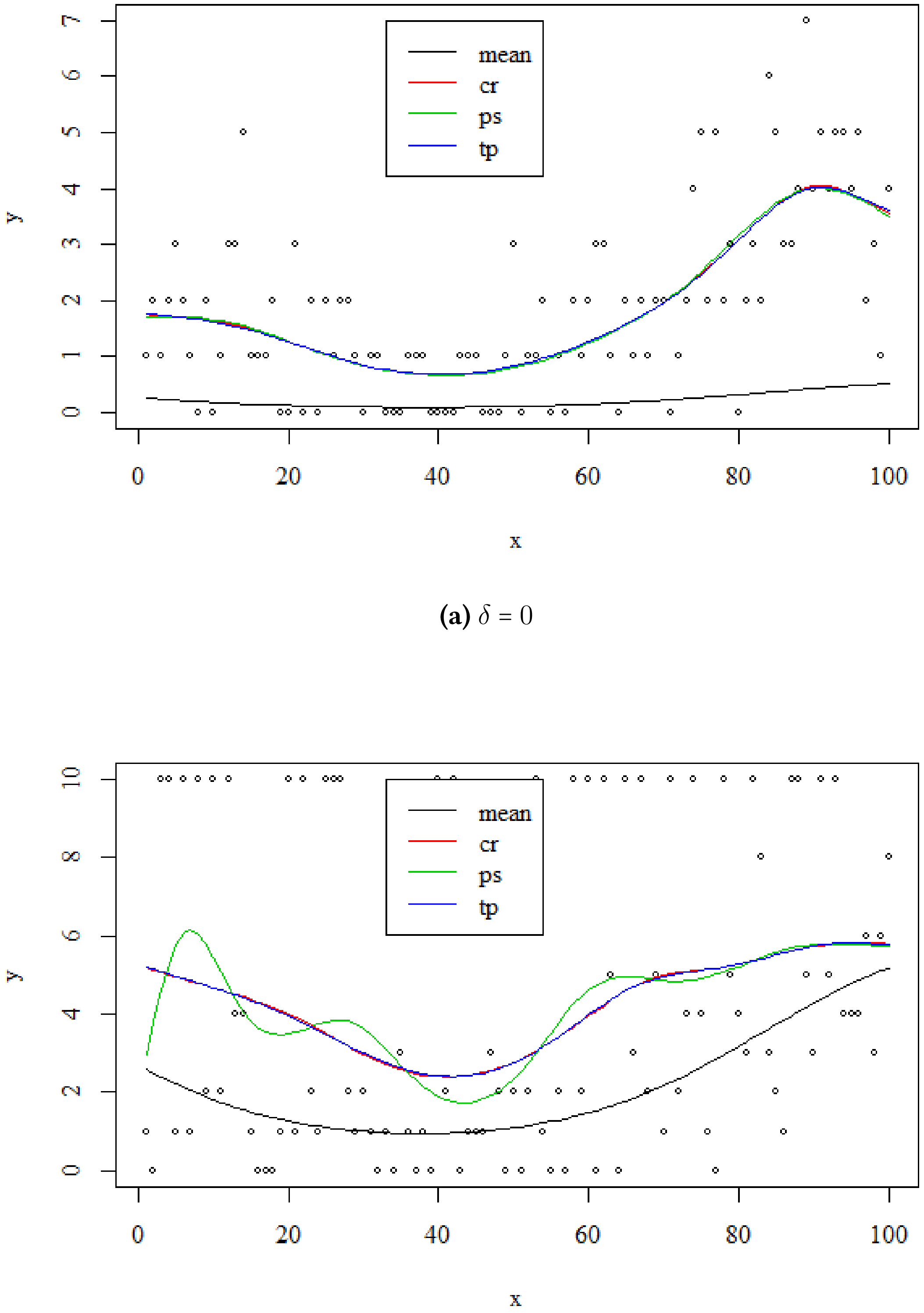


(a) $\delta = 0$



(b) $\delta = 0.2$

Figure 5.3. *GAM models fitted using cubic regression spline, p-spline and thin-plate spline for data with binomial response which is generated using Scenario 2*



(a) $\delta = 0$

(b) $\delta = 0.2$

Figure 5.4. *GAM models fitted using cubic regression, p-spline and thin-plate spline for data with binomial response which is generated using Scenario 4*

Predictions of the fitted models and the true mean are (black-colored line). From these figures, one can see that all the three methods are pulled away from the true mean at positions where outliers exist which makes clear that existence of abnormal observations somehow inflates GAM model deviances. Particularly, p-spline shows more fluctuations because of outliers.

5.2 Performance of Models in the Presence of Multicollinearity

Before proceeding with modeling the datasets, the presence of multicollinearity was checked using VIF method. From Table 5.3, it can be seen that it is less likely the generated data to suffer from multicollinearity when ρ is less than 0.9 which is why greater values are considered in the simulation.

Table 5.3. *VIF values of a simulated sample for* $n = 100$ *and* $\sigma^2 = 1$

ρ	x_1	x_2	x_3	x_4
0.8	2.27	2.28	2.07	2.04
0.9	5.52	4.74	6.36	5.42
0.99	40.00	49.55	37.89	47.73
0.999	347.51	341.81	396.64	309.72

Summary of the experiments of fitting GAM models using the three smoothing spline bases to datasets of varying size and different degree of collinearity are provided in Table 5.4. In addition, performance measures of GLM model is provided. The results in all the cases show that the GAM models have resulted in a smaller mean deviance compared to GLM. Only in a very few situations ($\delta = 0.99$, 0.999 and $n = 50$), GLM scored few small deviances. Inexplicably, this result shows that penalized regression spline based GAM models are less prone to the effect of multicollinearity than GLMs.

When the case of multicollinearity is taken into consideration to compare GAMs fitted using the three penalized regression splines, cubic regression spline was found to be the dominant which in all the cases has produced smaller mean deviance. More than 40% of the times, model produced using cubic regression showed smaller deviance. This being the fact, the standard deviation of model deviances using cubic regression spline are found larger than the others which indicates that this method is less consistent than the others. As it was presented in section 5.1, in the presence of multicollinearity too, p-spline was found fit models which are more close to each other.

Despite the fact that an increase in error variance increases the model deviances,

Table 5.4. *Mean, standard deviation and proportion of the number of times a model resulted in a small deviance in the presence of multicollinearity*

		glm		cr		ps		tp		glm	cr	ps	tp
ρ	σ^2	MD	SD	MD	SD	MD	SD	MD	SD	P			
n = 50													
0.9	1	45.21	9.24	31.67	11.88	35.45	10.63	36.49	10.7	0	0.53	0.3	0.17
	9	406.91	83.17	285.33	107.28	319.11	95.77	328.41	96.28	0	0.54	0.31	0.15
	25	1130.3	231.02	792.87	297.69	886.33	265.98	912.19	267.45	0	0.54	0.31	0.15
	100	4521.2	924.08	3180.44	1205.32	3547.06	1064.18	3649.8	1070.79	0	0.54	0.32	0.14
0.99	1	45.21	9.21	33.33	12.37	37.66	10.93	38.37	10.94	0	0.5	0.28	0.22
	9	406.88	82.87	302.37	113.29	340.66	99.99	345.42	98.55	0.01	0.48	0.28	0.23
	25	1130.23	230.19	841.23	314.93	945.84	278.01	959.69	273.66	0	0.47	0.29	0.24
	100	4520.93	920.76	3365.54	1266.69	3787.68	1112.85	3838.7	1094.66	0.01	0.48	0.29	0.23
0.999	1	45.21	9.2	37.03	13.09	40.39	11.04	41.19	10.71	0.02	0.39	0.28	0.31
	9	406.86	82.78	334	117.85	364.68	100.01	371.45	96.82	0.06	0.38	0.28	0.29
	25	1130.18	229.93	929.73	327.45	1010.37	276.18	1032.82	269.24	0.06	0.38	0.27	0.29
	100	4520.7	919.74	3712.1	1312.18	4043.48	1105.66	4130	1076.04	0.07	0.39	0.27	0.28
n = 100													
0.9	1	95.26	12.96	83.34	14.54	86.24	14.38	87.1	14.61	0	0.48	0.3	0.22
	9	857.33	116.66	750.05	130.85	776.19	129.43	783.87	131.59	0	0.49	0.3	0.22
	25	2381.48	324.06	2083.76	363.55	2156.07	359.54	2177.41	365.53	0	0.49	0.31	0.2
	100	9525.94	1296.25	8338.95	1458.85	8625.28	1440.24	8709.62	1462.11	0	0.49	0.31	0.2
0.99	1	95.27	12.97	85.47	15.29	88.22	14.43	88.86	14.49	0	0.44	0.31	0.24
	9	857.46	116.76	773.8	141.55	794.31	129.92	799.77	130.42	0	0.43	0.32	0.24
	25	2381.83	324.34	2149.52	391.9	2206.03	361.03	2221.56	362.21	0	0.42	0.34	0.23
	100	9527.32	1297.37	8596.86	1563.1	8825.63	1445.27	8886.22	1449.08	0	0.42	0.36	0.22
0.999	1	95.28	12.98	88.03	16.1	90.24	14.47	91.54	14.24	0	0.38	0.29	0.33
	9	857.51	116.8	793.08	143.84	814.13	130.44	824.05	128.14	0	0.37	0.31	0.32
	25	2381.97	324.43	2207.08	403.65	2264.32	365.1	2289.04	355.87	0	0.36	0.29	0.34
	100	9527.89	1297.72	8819.16	1604.72	9051.43	1457.74	9155.99	1423.1	0	0.36	0.28	0.36
n = 500													
0.9	1	492.36	34.63	481.22	35.11	482.4	34.84	483.8	34.95	0	0.42	0.36	0.22
	9	4431.24	311.68	4331.35	316.51	4341.6	313.57	4354.23	314.56	0	0.42	0.36	0.22
	25	12309	865.78	12031.52	879.2	12059.99	871.03	12095.1	873.79	0	0.42	0.37	0.21
	100	49235.99	3463.12	48127.12	3519	48239.97	3484.13	48380.38	3495.15	0	0.43	0.36	0.21
0.99	1	492.38	34.63	482.38	35.54	484.04	35.27	485.67	35.51	0	0.4	0.33	0.27
	9	4431.39	311.64	4341.66	320.26	4356.34	317.39	4371.04	319.61	0	0.41	0.36	0.24
	25	12309.43	865.68	12063.12	889.66	12100.94	881.63	12141.78	887.81	0	0.4	0.35	0.25
	100	49237.7	3462.72	48254.71	3563	48403.76	3526.53	48567.11	3551.24	0	0.43	0.38	0.2
0.999	1	492.38	34.63	483.68	35.59	485.31	35	488.52	35.47	0	0.39	0.3	0.31
	9	4431.44	311.64	4359.66	322.16	4368.11	314.97	4396.72	319.29	0	0.36	0.35	0.29
	25	12309.57	865.67	12109.08	894.94	12132.98	874.73	12213.11	886.9	0	0.37	0.33	0.3
	100	49238.27	3462.68	48436.13	3577.02	48534.62	3500.11	48852.45	3547.61	0	0.36	0.35	0.29

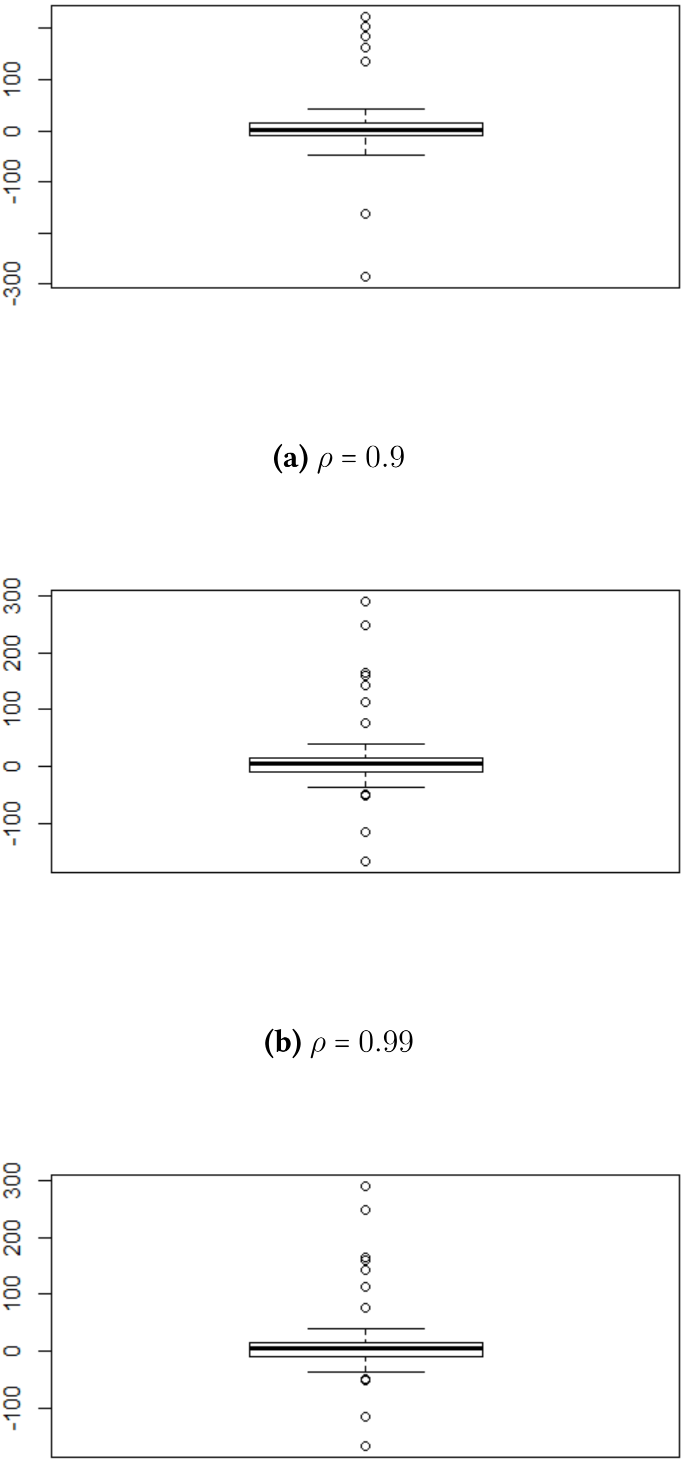
MD: Mean deviance; SD: Standard Deviation

P: proportion of the number of times a model produced smaller deviance

the proportion of ”smaller deviance” of a model for different values of σ^2 was found to be similar for a given sample size and correlations coefficient.

5.3 Performance of Models in the Presence of both Outliers and Multicollinearity

In addition to checking for the presence of multicollinearity as in section 5.2, the existence of abnormal observations was also checked using boxplot method. Figure 5.5 shows the existence of abnormal observations in one random sample of the simulation. It assures that there exists outliers as desired.



(a) $\rho = 0.9$

(b) $\rho = 0.99$

(c) $\rho = 0.999$

Figure 5.5. *Boxplot for checking the presence of outliers for one sample of size* $n = 100$ *with* $\delta = 0.1$

In this section, a data with multi-collinear covariates and outlier containing response variable were used to evaluate the performance of the penalized regression spline smoothers. All the settings used in section 5.2 are carried over to this section except that outliers ($\delta = 0.1$) are introduced to the response variable, y.

Table 5.5. *Mean, standard deviation and proportion of the number of times a model resulted in a small deviance in the presence of multicollinearity and outliers*

		glm		**cr**		ps		**tp**		**glm**	**cr**	**ps**	**tp**
ρ	σ^2	MD	SD	MD	SD	MD	SD	MD	SD	P			
n = 50													
0.9	1	517272	281082.1	208930.1	152117	184011.1	141645.6	213666	159647.1	0	0.35	0.4	0.25
	9	531121.2	287580.2	222386.2	165438.5	195637.5	149765	224407.3	166229.7	0	0.35	0.4	0.25
	25	559605.7	303409.1	242757.9	181462.2	216131.4	162565.4	248508.6	184494.8	0	0.35	0.42	0.23
	100	694846.2	383903.6	337406.8	256052.3	325533.8	240610.5	360383.4	263712.8	0	0.4	0.37	0.23
0.99	1												
	9	606551.2	322717.6	257337.5	198589.6	231280.8	180014.7	2 56090.3	207382.5	0	0.3	0.39	0.3
	25												
	100	768214.1	419584.9	373753	281735.7	357824.5	271167.6	388436.7	291616.8	0.01	0.32	0.35	0.32
0.999	1	601828.5	319564.3	291836.2	214038.6	253975.4	202350.5	335633.1	253379.2	0.03	0.26	0.43	0.28
	9	614941.4	325758.4	302522	224347.3	264924.2	209010	350831.8	266307.5	0.04	0.27	0.4	0.29
	25	642683.7	341598	326311	243205.3	286354.5	224546	372788.7	281339.9	0.04	0.26	0.39	0.3
	100	776043.2	422774.5	430559.1	323636.1	398028.1	303840.9	492275.8	361984.3	0.04	0.28	0.38	0.3
n = 100													
0.9	1	1039693	435334.4	690920.8	276243.3	571711.2	242400.5	615236.8	255431.9	0	0.16	0.58	0.26
	9	1073763	452668.3	718051.2	290205.4	599539.9	261785.1	642825.2	270443.2	0	0.18	0.58	0.24
	25	1138978	483395.7	773008.3	319731.7	656419.8	294378.1	701056.4	302359.5	0	0.21	0.55	0.25
	100	1438268	618015.7	1030875	444948.5	920297	434962.8	972294.6	450249.8	0	0.27	0.5	0.24
0.99	1	1187865	492116.3	788311.3	320400.5	639734.5	278077.5	701328.9	307867.4	0	0.16	0.57	0.27
	9	1221594	507637.6	816830.7	337033.1	674028.7	301376.7	729744.5	328405.1	0	0.17	0.53	0.3
	25	1286474	536655.2	871105.9	363237.3	730138.3	328053.1	784078.2	366654.8	0	0.2	0.5	0.3
	100	1584954	667887.4	1125960	485262.9	996720.1	461848.9	1054633	492437.2	0	0.23	0.49	0.28
	1	1204589	497665.9	825424.8	336694	685864.1	319013.9	795739	376898.9	0	0.15	0.58	0.28
	9	1238168	512657	861077.3	365384	714995.2	335129.9	825148.2	389829.5	0	0.16	0.56	0.28
	25	1302898	541129.8	917778.7	393644.5	772964.4	364406.1	884510.2	415520.8	0	0.18	0.57	0.25
	100	1601014	671098.7	1180302	525893.6	1049903	494016.7	1175404	543748.3	0	0.21	0.52	0.27
n = 500													
0.9	1	5605144	994099.6	5141729	849183	4651736	765800.4	4913665	798060.2	0	0.03	0.82	0.15
	9	5764672	1030860	5298754	889697.4	4809807	801570.8	5072344	837334.2	0	0.03	0.8	0.17
	25	6083831	1098075	5612594	959166.2	5127225	863057.7	5389632	908623	0	0.03	0.8	0.18
	100	7580117	1392780	7083469	1256762	6608243	1143752	6870437	1199873	0	0.06	0.75	0.19
0.99	1	6366311	1121806	5880103	964508.1	5271732	849479.1	5620942	910032.7	0	0.02	0.83	0.15
	9	6528053	1159690	6039524	1007990	5437500	891081	5785616	955548.7	0	0.02	0.82	0.16
	25	6849422	1228252	6354631	1080549	5752976	953712.3	6106812	1027134	0	0.03	0.83	0.14
	100	8351220	1526513	7826701	1378739	7249060	1247056	7595846	1322976	0	0.05	0.79	0.15
0.999	1	6443169	1132450	5971687	979141.2	5421672	891407.2	5809956	951467.4	0	0.02	0.84	0.13
	9	6605561	1170367	6132023	1022491	5584804	917395.5	5968423	989574.7	0	0.02	0.85	0.12
	25	6927581	1239039	6449151	1096999	5906202	986484.6	6291122	1063839	0	0.03	0.84	0.14
	100	8430996	1537732	7928056	1403790	7407992	1274173	7796092	1383570	0	0.04	0.8	0.16

MD: Mean deviance; SD: Standard Deviation

P: proportion of the number of times a model produced smaller deviance

Table 5.5 provides mean, standard deviation and proportion of smaller deviances of models fitted using glm and the three smoothing splines for situations where both outliers and multicollinearity are present in the dataset.

The results show that in all the situations, p-spline produced a smaller mean deviance. Moreover, this method was found to be more consistent in the performance of the model fitted for different samples. Considering the proportions of how many times a model produced a smaller deviance, the most striking result was emerged from p-spline method. It is remarkable that with the increase of sample size the number of times the p-spline method produced a smaller deviance has increased drastically. To illustrate this consider the combinations ($n = 50$, $\rho = 0.9$) and ($n = 500$, $\rho = 0.999$): in the first case approximately 40% of the times a smaller deviance was attained using p-spline whereas, in the later case p-spline produced smaller deviance more than 80% of the times. On the other hand, cubic regression loses its dominance of producing smaller mean deviance when outliers are included to the data. More essentially, it is demonstrated that the mean deviance of all the models increased with the increase of the degree of linear relationship given n and σ^2 are kept fixed.

6. CONCLUSION

In this thesis, three penalized regression spline smoothers are evaluated in fitting generalized additive model for a simulated data which contain outliers in the response variable or have predictor variables with linear relationship among them. The first is cubic spline, a curve made up of sections of cubic polynomials which are joined together and are continuous up to second derivatives. Another is the p-spline, which is fitted using b-spline with penalty. With cubic or p-splines, on top of defining the basis functions, knots have to be specified in order to operate the fitting procedure. The third is the thin-plate spline which avoids selection of basis functions and specifying knots positions.

In this thesis, three main studies have been performed; the first is a comparative study to find out a better performing method based on their model deviance when outliers are included in response variable. In general, in cases where outliers are present, p-spline is found to perform least best.

The second part of the study is performed under the existence of multicollinearity. In this study, model fitted using the smoothing splines of interest are compared with that of generalized linear model. Here, cubic regression was found to produce a better models.

In the last part of the study where existence of both multicollinearity and outliers were taken into consideration, models fitted using p-spline resulted in a smaller mean deviance.

By way of conclusion, the results seem to demonstrate that penalized smoothing splines could be used instead of generalized linear models when multicollinearity and outliers are present in a dataset.

REFERENCES

[1] J. A. Nelder and R. W. M. Wedderburnn, "Generalized linear models," *Journal of the Royal Statistical Society*, vol. Series A 135, no. 3.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. John Wiley & Sons, Inc., 2012.

[3] p. McCullagh and J. Nelder, *Generalized Additive Models*, 2nd ed. Chapman & Hall/CRC, 1989.

[4] S. N. Wood, *Generalized Additive Models; An Introduction with R.* Chapman & Hall/CRC, 2006.

[5] S. He, "Generalized additive models for data with concurvity: statistical issues and a novel model fitting approach," Ph.D. dissertation, University of Pittsburgh, 2004.

[6] L. Keele, *Semiparametric Regression for the Social Sciences.* John Wiley & Sons Ltd, 2008.

[7] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models.* Chapman & Hall, 1990.

[8] H. Wu and J.-T. Zhang, *Nonparametric Regression Methods for Longitudinal Data Analysis; Mixed-Effects Modeling Approaches.* John Wiley & Sons, Inc., 2006.

[9] T. Hastie, *gam: Generalized Additive Models*, 2016, r package version 1.14. [Online]. Available: https://CRAN.R-project.org/package=gam

[10] J. J. Faraway, *Extending the Linear Model with R - Generalized Linear, Mixed Effects and Nonparametric Regression Models.* taylor & Francis Group, LLC, 2006.

[11] A. Alimadad and M. Salibian-Barrera, "An outlier-robust fit for generalized additive models with applications to disease outbreak detection," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 719–731, 2011.

[12] R. K. Wong, F. Yao, and T. C. Lee, "Robust estimation for generalized additive models," *Journal of Computational and Graphical Statistics*, vol. 23, no. 1, pp. 270–289, 2014.

[13] A. Buja, T. Hastie, and R. Tibshirani, "Linear smoothers and additive models," *The Annals of Statistics*, vol. 17, no. 2, 1989.

[14] S. Amodio, M. Aria, and A. D'Ambrosio, "On concurvity in nonlinear and nonparametric regression models," *Statistica*, no. 1, 2014.

[15] G. Rodriguez, "Smoothing and non-parametric regression," 2001. [Online]. Available: http://data.princeton.edu/eco572/smoothing.pdf

[16] S. Yakowitz and F. Szidarovszky, *An Introduction to Numerical Computation.* Macmillan Publishing Company, 1990.

[17] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* Chapman & Hall, 1994.

[18] D. Ruppert, M. P. Wand, and R. J. Caroll, *Semiparametric Regression.* Cambridge

University Press Cambridge, 2003.

[19] W. Shen, *An Introduction to Numerical Computation.* World Scientific, 2015.

[20] G. D. Knott, *Interpolating Cubic Splines.* Birkhauser Boston, 2000.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer Science+Business Media, LLC, 2009.

[22] C. Gu, *Smoothing Spline ANOVA Models.* Springer-Verlag New York, Inc, 2002.

[23] C. de Boor, *A Practical Guide to Splines.* Springer-Verlag New York, Inc., 1978.

[24] P. H. C. Eilers and B. D. Marx, "Flexible smoothing with b-splines and penalties," *Statistical Science*, vol. 11, no. 2, 1996.

[25] D. Eberly, "Thine plate splines," *Geometric Tools, LLC*, 1996.

[26] S. N. Wood, "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 65, no. 1, pp. 95–114, 2003.

[27] A. Agresti, *Foundations of Linear and Generalized Linear Models.* John Wiley & Sons, Inc., 2015.

[28] S. Wood and M. S. Wood, "The mgcv package," *www. r-project. org*, 2007.

[29] F. O'sullivan, B. S. Yandell, and W. J. Raynor Jr, "Automatic smoothing of regression functions in generalized linear models," *Journal of the American Statistical Association*, vol. 81, no. 393, pp. 96–103, 1986.

[30] G. Wahba, *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics: Philadelphia, 1990.

[31] R. H. Myers, D. C. Montgomery, and G. G. Vining, *Generalized Linear Models with Applications in Engineering and the Sciences*, 2nd ed. John Wiley & Sons, Inc, 2010.

[32] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. Springer, 2002.

[33] G. C. McDonald and D. I. Galarneau, "A monte carlo evaluation of some ridge-type estimators," *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 407–416, 1975.