Nikola Rogler

# Generalized Additive Models and their Use for Gas Flow Data

**MASTERARBEIT**

zur Erlangung des akademischen Grades einer/s
Diplom-Ingenieur/in

Masterstudium Finanz- und Versicherungsmathematik


Technische Universität Graz

Betreuer/in:
Ao.Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig Friedl

Institut für Statistik

Graz, im April 2013

# Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ..... .....

(Unterschrift)

# Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

..... .....

date (signature)

# Contents

# Preface

The most astonishing experience while working with generalized additive models was that they are very similar to generalized linear models but offer the possibility to estimate nonlinear relationships in a very simple and fascinating way. For this interesting topic and all the support and feedback I got while writing this thesis I want to especially thank Prof. Herwig Friedl.

The possibility to go to Linz to participate in a workshop about P-splines held by Brian Marx was a great opportunity for me for which I want to thank Joanneum Research and especially Dr. Franz Prettenthaler. During this workshop I had the chance to see how P-splines work and how they can be used for generalized additive models.

Finally, I want to thank Arno Kimeswenger for supporting me in my work and helping me during the process of correction. In addition, I want to thank the countless number of people who supported me in their own way or helped me by sharing their thoughts with me on this topic. All of them I send my thanks.

# 1. Introduction

The motivation of this work was to estimate the mean of the daily maximum gas flow provided by the Open Grid Europe GmbH (OGE) depending on the mean daily temperature, see Friedl, Mirkov, and Steinkamp (2012). The data contains gas flow maxima per day of one knot of the German gas transmission network from January 2004 to the end of June in 2009, in total 2008 data points. In addition, the daily mean temperature is also included in the dataset. Later on we will see that the gas flow is dependent on the mean temperature but in a nonlinear sigmoid shaped way.

To estimate the nonlinear relationship between mean temperature and daily maximal gas flow we have two options:

- a nonlinear mean model,
- a generalized additive model (GAM).

The first option is addressed in Friedl et al. (2012) by considering sigmoidal models and assuming normally distributed responses, while this work concentrates on generalized additive models.

The advantage of GAMs lies in the fact that they can describe any nonlinear relationship very well but can also be reduced to generalized linear models, which makes them easy to handle. Moreover, since we will be looking at daily maxima, the normal distribution might not be appropriate. Another advantage of GAMs is the possibility to choose different distributions from the exponential family. For example, a gamma distribution could be the better option in this case than a normal distribution. More information on GAMs can be found for example in Wood (2006a), Eilers and Marx (1996) or Hastie and Tibshirani (1990).

In addition, the issue of how to estimate the mean at temperatures below $-10^{\circ}$ Celsius is addressed. This is a major issue for the OGE because they have to guarantee sufficient gas flow as far as $-12^{\circ}$C to $-16^{\circ}$C depending on the area in Germany. Therefore one aspect of this work will concentrate on how to estimate the mean and how to construct confidence and prediction intervals.

For this work the software R, see R Development Core Team (2011), is used, especially the package `mgcv`, which is discussed in Wood (2006a). The plots in this work were produced with the R package `ggplot2`, whose options and functions are discussed in Wickham (2009). In addition, the code to estimate the model parameters and the code to produce the plots of this work is added in the various chapters or in the appendix.

After this part we start with an introduction to linear models in Chapter 2. There we assume normally distributed responses and a linear relationship between the mean and the predictor variable. Thereafter, we continue with generalized linear models in

Chapter 3, where a function of the mean is described by a linear predictor and the distribution of the response variable is a member of the exponential family. In a next step we look at possibilities to estimate a nonlinear relationship with additive models assuming a normal distribution, see Chapter 4. Finally, we change the assumption of a normal distribution to any other distribution from the exponential family and consider generalized additive models in Chapter 5. The theory of these chapters is then applied to the gas flow data in Chapter 6. Finally, in Chapter 7 the results of the previous chapters are summarized.

# 2. Linear Models

The purpose of this first chapter is to give a brief introduction to Linear Models, so that later on more complex models can be established. For further information see for example Wood (2006a).

In a linear model

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

the response variable $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is described by $p$ predictor variables. At each observation $i$ the vector of the explanatory variables is $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip-1})^T$ for $i = 1, \ldots, n$, which summarizes to the model matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$. The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$ stands for the parameters that need to be estimated, while $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\,\boldsymbol{I}_n)$ is normally distributed with zero mean and variance matrix $\sigma^2\,\boldsymbol{I}_n$. Therefore, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\,\boldsymbol{\beta}, \sigma^2\,\boldsymbol{I}_n)$ is normally distributed with $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X}\,\boldsymbol{\beta}$. In addition, $n$ refers to the number of observations and $p-1$ to the number of covariates included.

In the following sections different issues regarding linear models will be discussed. First of all, the estimation of $\boldsymbol{\beta}$ is addressed.

## 2.1. Parameter Estimation

### 2.1.1. Least Squares Estimation

One common approach to estimate the parameters in model equation (2.1) is by least squares estimation. In this case, $\hat{\boldsymbol{\beta}}$ is solution of the minimization of

$$\mathrm{SSE}(\boldsymbol{\beta}) = \| \boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta} \|^2 = \sum_{i=1}^{n} (y_i - \mu_i)^2,$$

where $\mu_i = \boldsymbol{x}_i^T\,\boldsymbol{\beta}$.

To get an estimator of $\boldsymbol{\beta}$ we take a look at the first derivative of $\mathrm{SSE}(\boldsymbol{\beta})$

$$\frac{\partial\,\mathrm{SSE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\,\boldsymbol{X}^T\,(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})\,.$$

Setting the first derivative to zero leads to

$$\boldsymbol{X}^T\,\boldsymbol{y} = \boldsymbol{X}^T\,\boldsymbol{X}\,\hat{\boldsymbol{\beta}}\,.$$

If the inverse exists, the estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}\,. \tag{2.2}$$

Since $\hat{\boldsymbol{\beta}}$ is a linear combination of $\boldsymbol{y}$, the assumption that $\boldsymbol{y}$ is normally distributed implies that $\hat{\boldsymbol{\beta}}$ is normally distributed too. Furthermore, it can be shown that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator with variance $\sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$, i.e.

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \mathbb{E}(\boldsymbol{y}) = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{X}\, \boldsymbol{\beta} = \boldsymbol{\beta}, \\
\operatorname{var}(\hat{\boldsymbol{\beta}}) &= \operatorname{var}\left(\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}\right) = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \operatorname{var}(\boldsymbol{y})\; \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \\
&= \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}.
\end{aligned}$$

Together we get

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right). \tag{2.3}$$

An estimator of $\boldsymbol{\mu}$ is constructed as

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\, \hat{\boldsymbol{\beta}} = \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} = \boldsymbol{H}\, \boldsymbol{y},$$

where $\boldsymbol{H} = \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T$ is the so called hat matrix. Before some properties of the hat matrix like symmetry, idempotence and projection are discussed in the course of this chapter, the mean and variance of $\hat{\boldsymbol{\mu}}$ are illustrated. Since $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, it follows that $\hat{\boldsymbol{\mu}}$ is also unbiased, since

$$\begin{aligned}
\mathbb{E}\left(\hat{\boldsymbol{\mu}}\right) &= \boldsymbol{X}\, \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\, \boldsymbol{\beta} = \boldsymbol{\mu}, \\
\operatorname{var}\left(\hat{\boldsymbol{\mu}}\right) &= \boldsymbol{X} \operatorname{var}(\hat{\boldsymbol{\beta}})\, \boldsymbol{X}^T = \sigma^2\, \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T = \sigma^2\, \boldsymbol{H}\,.
\end{aligned}$$

Three properties of the hat matrix will be used in the sequel:

- $\boldsymbol{H} = \boldsymbol{H}^T$ (symmetry)

$$\boldsymbol{H}^T = \left(\boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\right)^T = \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T = \boldsymbol{H},$$

- $\boldsymbol{H}\, \boldsymbol{H} = \boldsymbol{H}$ (idempotence)

$$\begin{aligned}
\boldsymbol{H}\, \boldsymbol{H} &= \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \\
&= \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T = \boldsymbol{H},
\end{aligned}$$

- $\boldsymbol{H}\, \boldsymbol{X} = \boldsymbol{X}$ (projection)

$$\boldsymbol{H}\, \boldsymbol{X} = \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{X}\,.$$

The residuals $\boldsymbol{r}$ describe the difference between the observed response values $\boldsymbol{y}$ and the fitted values $\hat{\boldsymbol{\mu}}$

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - \boldsymbol{H}\,\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}\,.$$

As a consequence of the normal distribution of $\boldsymbol{y}$, $\boldsymbol{r}$ is also normally distributed with mean

$$\mathbb{E}\,(\boldsymbol{r}) = (\boldsymbol{I} - \boldsymbol{H})\,\mathbb{E}\,(\boldsymbol{y}) = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{X}\,\boldsymbol{\beta} = \boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{H}\,\boldsymbol{X}\,\boldsymbol{\beta} = \boldsymbol{0}$$

and variance

$$\mathrm{var}(\boldsymbol{r}) = (\boldsymbol{I} - \boldsymbol{H})\,\mathrm{var}(\boldsymbol{y})\,(\boldsymbol{I} - \boldsymbol{H})^T = \sigma^2\,(\boldsymbol{I} - \boldsymbol{H})\,.$$

The last result is due to the fact that $(\boldsymbol{I} - \boldsymbol{H})$ is idempotent, since

$$(\boldsymbol{I} - \boldsymbol{H})\,(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - 2\,\boldsymbol{H} + \boldsymbol{H}\,\boldsymbol{H} = \boldsymbol{I} - 2\,\boldsymbol{H} + \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{H},$$

because $\boldsymbol{H}$ is idempotent.

### 2.1.2. Maximum Likelihood Estimation

Another possible approach to estimate the model parameters is to consider maximum likelihood estimation. Under the assumption of a normal distribution the likelihood function is given by

$$L\left(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})\right\},$$

while the log-likelihood is

$$\begin{aligned} l(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2) &= \log L\left(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\,(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})^T\,(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})\,. \end{aligned}$$

To obtain an estimate for $\boldsymbol{\beta}$ and $\sigma^2$ the partial derivatives are calculated

$$\begin{aligned} \frac{\partial l}{\partial\,\boldsymbol{\beta}} &= \frac{1}{\sigma^2}\,\boldsymbol{X}^T\,(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta})\,, \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\|\,\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}\,\|^2. \end{aligned}$$

Setting these first derivatives to zero, we get the estimators

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{y}, \\ \hat{\sigma}^2 &= \frac{1}{n}\|\,\boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}}\,\|^2. \end{aligned}$$

One may notice that the estimator $\hat{\boldsymbol{\beta}}$ equals the least squares estimator in (2.2). In fact in case of normally distributed responses the least squares and the maximum likelihood estimation yield the same results.

## 2.2. Hypothesis Testing

### 2.2.1. t Test

While the last section concentrated on the estimation of the model parameters, this section answers the question if every single covariate $\boldsymbol{x}_j$ is needed, where $\boldsymbol{x}_j$ denotes the $j$-th column of of the model matrix $\boldsymbol{X}$. Therefore, we test the hypothesis

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0 \tag{2.4}$$

for any $j = 1, \ldots, p-1$.

Since the normal distribution of $\hat{\boldsymbol{\beta}}$ was shown in (2.3) , it follows that

$$\frac{1}{\sigma}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{1/2} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_p\right), \tag{2.5}$$

where $\boldsymbol{\beta}$ is the true parameter and $\hat{\boldsymbol{\beta}}$ its estimator. Now we show that for normal responses, $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ or $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}^T\,\boldsymbol{r} = \mathrm{SSE}(\hat{\boldsymbol{\beta}})$ are independent. On that account we write $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ in terms of $\boldsymbol{\varepsilon}$ and get

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{y} = \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{\beta} + \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{\varepsilon},\\
\boldsymbol{r} &= \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\left(\boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \boldsymbol{X}\,\boldsymbol{\beta} - \boldsymbol{X}\,\boldsymbol{\beta} + \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{\varepsilon}\\
&= \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{\varepsilon}\,.
\end{aligned}$$

By analysing the covariance between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ one gets

$$\begin{aligned}
\mathrm{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r}) &= \mathrm{cov}\left(\boldsymbol{\beta} + \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{\varepsilon}, \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{\varepsilon}\right)\\
&= \mathrm{cov}\left(\left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{\varepsilon}, \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{\varepsilon}\right)\\
&= \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\mathrm{cov}\left(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}\right)\left(\boldsymbol{I} - \boldsymbol{H}\right)\\
&= \sigma^2\left(\left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T - \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{H}\right)\\
&= \sigma^2\left(\left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T - \left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\,\boldsymbol{X}\left(\boldsymbol{X}^T\,\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)\\
&= \boldsymbol{0}\,.
\end{aligned} \tag{2.6}$$

In the next step we point out that $\mathrm{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$. To show this we use

$$\begin{aligned}
\boldsymbol{\varepsilon}^T\,\boldsymbol{\varepsilon} &= \left(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\beta}\right)\\
&= \left(\boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}} + \boldsymbol{X}\,\hat{\boldsymbol{\beta}} - \boldsymbol{X}\,\boldsymbol{\beta}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\,\hat{\boldsymbol{\beta}} + \boldsymbol{X}\,\hat{\boldsymbol{\beta}} - \boldsymbol{X}\,\boldsymbol{\beta}\right)\\
&= \left(\boldsymbol{r} + \boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right)^T\left(\boldsymbol{r} + \boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right).
\end{aligned}$$

Since $\boldsymbol{r} = \left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y}$ and $\boldsymbol{H}\,\boldsymbol{X} = \boldsymbol{X}$, it follows that the mixed term in the above equation is zero, i.e.

$$\boldsymbol{r}^T\,\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \boldsymbol{0}\,.$$

Therefore,

$$\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \boldsymbol{r}^T \boldsymbol{r} + \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{2.7}$$

Because $\boldsymbol{\varepsilon}$ is normally distributed with zero mean and variance $\sigma^2 \boldsymbol{I}_n$, it follows that

$$\frac{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \varepsilon_i^2 \sim \chi_n^2.$$

In the same way, the normal distribution of $\hat{\boldsymbol{\beta}}$ (see (2.5)) ensures that

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \frac{1}{\sigma^2} \sim \chi_p^2.$$

Taking the moment-generating function of the $\chi^2$-distribution with $n$ degrees of freedom

$$M(t) = (1 - 2t)^{-n/2}, \quad t < \frac{1}{2}$$

and the independence of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ into account, we get from equation (2.7) that

$$(1 - 2t)^{-n/2} = \mathbb{E}\left(e^{t\, \boldsymbol{r}^T \boldsymbol{r} / \sigma^2}\right) (1 - 2t)^{-p/2}.$$

Therefore,

$$\mathbb{E}\left(e^{t\, \boldsymbol{r}^T \boldsymbol{r} / \sigma^2}\right) = (1 - 2t)^{-(n-p)/2}$$

and $\boldsymbol{r}^T \boldsymbol{r} / \sigma^2 = \frac{1}{\sigma^2} \operatorname{SSE}(\hat{\boldsymbol{\beta}}) \sim \chi_{n-p}^2$. Regarding that information, it follows that

$$\frac{1}{\sigma^2} \mathbb{E}\left(\operatorname{SSE}(\hat{\boldsymbol{\beta}})\right) = n - p,$$
$$\frac{1}{\sigma^4} \operatorname{var}\left(\operatorname{SSE}(\hat{\boldsymbol{\beta}})\right) = 2(n - p).$$

In this way we also found an unbiased estimator of $\sigma^2$, namely

$$\hat{\sigma}^2 = \frac{1}{n - p} \operatorname{SSE}(\hat{\boldsymbol{\beta}}).$$

As a result one is able to test the hypothesis (2.4) by computing the test statistic

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 d_{jj}}},$$

where $d_{jj}$ stands for the respective diagonal element of $\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$, and rejecting $H_0$ and thereby favouring $H_1$ if $|T| > t_{n-p,1-\alpha/2}$. Therefore, the two sided confidence interval for $\beta_j$ is given by

$$\left(\hat{\beta}_j - \hat{\sigma} d_{jj}^{1/2} t_{n-p,1-\alpha/2}, \hat{\beta}_j + \hat{\sigma} d_{jj}^{1/2} t_{n-p,1-\alpha/2}\right).$$

### 2.2.2. F Test

By partitioning the design matrix $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1, \boldsymbol{X}_2$, one can test if a group of covariates is unnecessary instead of testing for one parameter $\beta_j$. Therefore, we rewrite the model as

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\,\boldsymbol{\beta}_1 + \boldsymbol{X}_2\,\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ describe the disjoint subsets of $\boldsymbol{\beta}$ which correspond to the $n \times q$ matrix $\boldsymbol{X}_1$ and the $n \times (p-q)$ matrix $\boldsymbol{X}_2$ $(p > q)$.

In the following, we test the hypothesis that

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0} \text{ vs. } H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{0},$$

where the alternative hypothesis $H_1$ states that any components in $\boldsymbol{\beta}_2$ are nonzero. In other words we want to know if the model $\boldsymbol{y} = \boldsymbol{X}_1\,\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ suffices.

Therefore, we look at the orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{X}_1$

$$\hat{\boldsymbol{\mu}}_1 = \boldsymbol{X}_1 \left(\boldsymbol{X}_1^T\,\boldsymbol{X}_1\right)^{-1} \boldsymbol{X}_1^T\,\boldsymbol{y}\,.$$

The residuals $\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1$ can be divided in two orthogonal parts

$$\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1 = (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1)\,.$$

Because of the Pythagorean theorem it follows that

$$(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1)^T\,(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_1) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T\,(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1)^T\,(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1)$$

or equivalently

$$\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) = \mathrm{SSE}(\hat{\boldsymbol{\beta}}) + \left(\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathrm{SSE}(\hat{\boldsymbol{\beta}})\right).$$

From the previous subsection we already know that $\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) \sim \sigma^2\chi^2_{n-q}$ and $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{n-p}$. In consequence of the independence of $\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathrm{SSE}(\hat{\boldsymbol{\beta}})$ and $\mathrm{SSE}(\hat{\boldsymbol{\beta}})$ it results that

$$\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathrm{SSE}(\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{p-q}.$$

Therefore, we know that

$$F = \frac{\left(\mathrm{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathrm{SSE}(\hat{\boldsymbol{\beta}})\right)/(p-q)}{\mathrm{SSE}(\hat{\boldsymbol{\beta}})/(n-p)} \sim F_{p-q,n-p},$$

and the null hypothesis is rejected if $F > F_{p-q,n-p;1-\alpha}$.

## 2.3. Confidence and Prediction Intervals

If a new vector $\boldsymbol{x}^* = \left(1, x_1^*, \ldots, x_{p-1}^*\right)^T$ is available, we are interested where $\mu^*$ and $y^*$ will probably lie. Since $\mu^* = \boldsymbol{x}^{*T}\boldsymbol{\beta}$, this mean can be estimated by $\hat{\mu}^* = \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}$. From (2.3) we know that $\hat{\boldsymbol{\beta}}$ is normally distributed and since $\hat{\mu}^*$ is a linear combination of $\hat{\boldsymbol{\beta}}$ it is normally distributed, too. The mean and variance are given by

$$\mathbb{E}\left(\hat{\mu}^*\right) = \mathbb{E}\left(\boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{x}^{*T}\boldsymbol{\beta}, \tag{2.8}$$

$$\operatorname{var}\left(\hat{\mu}^*\right) = \operatorname{var}\left(\boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{x}^{*T}\operatorname{var}\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{x}^* = \sigma^2\,\boldsymbol{x}^{*T}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*, \tag{2.9}$$

respectively.

As a result of the normal distribution of $\hat{\mu}^*$ and since $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}$ are independent, see (2.6), we know that

$$\frac{\boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} - \boldsymbol{x}^{*T}\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2\,\boldsymbol{x}^{*T}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*}} \sim t_{n-p}.$$

Therefore, the two sided $(1-\alpha)$ confidence interval for the true parameter $\mu^*$ is given by

$$\left(\boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} - \sqrt{\hat{\sigma}^2 h^*}\,t_{n-p,1-\alpha/2}, \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} + \sqrt{\hat{\sigma}^2 h^*}\,t_{n-p,1-\alpha/2}\right),$$

where $h^* = \boldsymbol{x}^{*T}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*$.

Next a prediction interval for the new observation $y^*$ is deduced. Because of (2.8) and (2.9), we know that

$$\begin{aligned}
\mathbb{E}\left(y^* - \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}\right) &= \boldsymbol{x}^{*T}\boldsymbol{\beta} - \boldsymbol{x}^{*T}\boldsymbol{\beta} = 0,\\
\operatorname{var}\left(y^* - \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}\right) &= \sigma^2\left(1 + \boldsymbol{x}^{*T}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*\right) = \sigma^2\left(1 + h^*\right).
\end{aligned}$$

This yields to

$$\frac{y^* - \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\left(1 + \boldsymbol{x}^{*T}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}^*\right)}} \sim t_{n-p}.$$

In this way we get a two sided $(1-\alpha)$ prediction interval for $y^*$, that is

$$\left(\boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} - \sqrt{\hat{\sigma}^2\left(1+h^*\right)}\,t_{n-p,1-\alpha/2}, \boldsymbol{x}^{*T}\hat{\boldsymbol{\beta}} + \sqrt{\hat{\sigma}^2\left(1+h^*\right)}\,t_{n-p,1-\alpha/2}\right).$$

# 3. Generalized Linear Models

Although linear models are very useful, they have some shortcomings. For instance, the assumption of a normal distribution is not always supported by the data. Especially count data is usually troublesome. Another drawback is the assumption of constant variance $\sigma^2$. An example for a violation of this assumption is a variance which increases proportional to the mean.

Therefore, we summarize in this chapter another model class which is not restricted by these assumptions. Further information on this model class can be found in Mc Cullagh and Nelder (1989), Friedl (2011) and Wood (2006a), on which this chapter is based upon.

A model that manages without the above mentioned restrictions is a generalized linear model given by

$$g\left(\mu_i\right) = \boldsymbol{x}_i^T \boldsymbol{\beta} = \eta_i, \quad i = 1, \ldots, n,$$

where $g(\cdot)$ is a smooth monotonic link function, $\boldsymbol{x}_i$ the $i$-th row of the design matrix $\boldsymbol{X}$, $\mu_i = \mathbb{E}\left(y_i\right)$, $\boldsymbol{\beta}$ the parameter vector and $\boldsymbol{\eta}$ the linear predictor. In addition, $\boldsymbol{y}$ is assumed to be independent distributed with a distribution from the exponential family. The definition of the exponential family can be found in the following section.

## 3.1. Exponential Family

Here we define the exponential family as used in GLMs. Further, we will discuss some important properties that will be useful for latter considerations.

**Definition.** *Exponential Family*
*A probability mass or density function of a random variable y belongs to the one parameter, linear exponential family with canonical parameter $\theta$, if it can be written as*

$$f\left(y|\theta\right) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c\left(y, \phi\right)\right\},$$

*where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are some known functions and $\phi$ is the dispersion parameter.*

Among others the normal distribution, the binomial and the Poisson distribution as well as the gamma distribution are members of the exponential family. In these cases the functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are defined as (or see Friedl (2011)):

- **Normal distribution:**

$$a(\phi) = \phi, \quad b(\theta) = \frac{\theta^2}{2}, \quad c(y, \phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi),$$

because for $y \sim \mathcal{N}(\mu, \sigma^2)$ the density function is given by

$$\begin{aligned} f(y, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\}, \end{aligned}$$

where $y \in \mathbb{R}$. Setting $\theta = \mu$ and $\phi = \sigma^2$, leads to the above definition of $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$.

- **Binomial distribution:**

$$a(\phi) = \phi = \frac{1}{k}, \quad b(\theta) = \log\left(\frac{1}{1-\pi}\right), \quad c(y, \phi) = \log\binom{\frac{1}{\phi}}{\frac{y}{\phi}},$$

since the probability mass function of the binomial distribution ($ky \sim \mathcal{B}(k, \pi)$) is defined as

$$\begin{aligned} f(y, k, \pi) &= \binom{k}{ky} \pi^{ky} (1-\pi)^{k-ky} \\ &= \exp\left\{\log\binom{k}{ky} + ky\log(\pi) + k(1-y)\log(1-\pi)\right\} \\ &= \exp\left\{\frac{y\log\left(\frac{\pi}{1-\pi}\right) - \log\left(\frac{1}{1-\pi}\right)}{1/k} + \log\binom{k}{ky}\right\}, \end{aligned}$$

where $y = 0, \frac{1}{k}, \frac{2}{k}, \ldots, 1$. Using $\theta = \log\left(\frac{\pi}{1-\pi}\right)$ and $\phi = 1/k$, the above functions follow.

- **Poisson distribution:**

$$a(\phi) = \phi = 1, \quad b(\theta) = \exp(\theta), \quad c(y, \phi) = -\log(y!),$$

because the probability mass function of a Poisson distribution is specified as

$$f(y, \mu) = \frac{\mu^y}{y!} \exp\{-\mu\} = \exp\{y\log(\mu) - \mu - \log(y!)\},$$

where $y = 0, 1, 2, \ldots$. Including $\theta = \log(\mu)$ and $\phi = 1$, the definition of the above functions ensues.

- **Gamma distribution:**

$$\begin{aligned} a(\phi) = \phi, \quad b(\theta) &= -\log(-\theta), \\ c(y, \phi) = \frac{1}{\phi}\log\left(\frac{1}{\phi}\right) + \left(\frac{1}{\phi} - 1\right)\log(y) &- \log\left(\Gamma\left(\frac{1}{\phi}\right)\right), \end{aligned}$$

since the gamma density function for $y \sim \mathcal{G}(\nu, \frac{\nu}{\mu})$ is

$$\begin{aligned} f(y,\mu,\nu) &= \exp\left\{-\frac{\nu}{\mu}y\right\}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\frac{1}{\Gamma(\mu)} \\ &= \exp\left\{-\frac{\nu}{\mu}y + \nu\log(\nu) - \nu\log(\mu) + (\nu-1)\log(y) - \log(\Gamma(\nu))\right\} \\ &= \exp\left\{\frac{-\frac{y}{\mu} + \log\left(\frac{1}{\mu}\right)}{1/\nu} + \nu\log(\nu) + (\nu-1)\log(y) - \log(\Gamma(\nu))\right\}, \end{aligned}$$

where $\mu, \nu, y > 0$. Assuming that $\theta = -\frac{1}{\mu}$ and $\phi = \frac{1}{\nu}$ the definitions of $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ follow.

The binomial and the Poisson distribution are often used in practice because of their usefulness regarding frequencies and count data, respectively. In addition, one might notice that in all examples above $a(\phi) = \phi$.

Next some properties of the score function, which is the derivative of the log-likelihood $l(y,\theta) = \log f(y,\theta)$, are pointed out. For any $f(y|\theta)$ from the exponential family, we have

$$\mathbb{E}\left(\frac{\partial l(y,\theta)}{\partial\theta}\right) = 0, \tag{3.1}$$

$$\mathbb{E}\left(-\frac{\partial^2 l(y,\theta)}{\partial\theta^2}\right) = \mathbb{E}\left(\left(\frac{\partial l(y,\theta)}{\partial\theta}\right)^2\right). \tag{3.2}$$

The following proof is taken from Friedl (2011).

$$\begin{aligned} \mathbb{E}\left(\frac{\partial l(y,\theta)}{\partial\theta}\right) &= \mathbb{E}\left(\frac{1}{f(y,\theta)}\frac{\partial f(y,\theta)}{\partial\theta}\right) = \int \frac{1}{f(y,\theta)}\frac{\partial f(y,\theta)}{\partial\theta} f(y,\theta)\,dy \\ &= \int \frac{\partial}{\partial\theta} f(y,\theta)\,dy = \frac{\partial}{\partial\theta}\int f(y,\theta)\,dy \end{aligned}$$

Since $\int f(y,\theta)\,dy = 1$, it follows that

$$\mathbb{E}\left(\frac{\partial l(y,\theta)}{\partial\theta}\right) = \frac{\partial}{\partial\theta}1 = 0.$$

To prove (3.2), we take a look at the second derivative

$$\frac{\partial^2 l(y,\theta)}{\partial\theta^2} = \frac{\partial^2 f(y,\theta)}{\partial\theta^2}\frac{1}{f(y,\theta)} - \frac{1}{f^2(y,\theta)}\left(\frac{\partial f(y,\theta)}{\partial\theta}\right)^2,$$

and its negative mean

$$\begin{aligned}
\mathbb{E}\left(-\frac{\partial^2 l(y,\theta)}{\partial\theta^2}\right) &= -\int \frac{\partial^2 l(y,\theta)}{\partial\theta^2} f(y,\theta)\,dy \\
&= -\int \frac{\partial^2 f(y,\theta)}{\partial\theta^2}\frac{1}{f(y,\theta)} f(y,\theta)\,dy + \int \frac{1}{f^2(y,\theta)}\left(\frac{\partial f(y,\theta)}{\partial\theta}\right)^2 f(y,\theta)\,dy \\
&= -\frac{\partial^2}{\partial\theta^2}\int f(y,\theta)\,dy + \int \left(\frac{\partial l(y,\theta)}{\partial\theta}\right)^2 f(y,\theta)\,dy \\
&= \int \left(\frac{\partial l(y,\theta)}{\partial\theta}\right)^2 f(y,\theta)\,dy = \mathbb{E}\left(\left(\frac{\partial l(y,\theta)}{\partial\theta}\right)^2\right).
\end{aligned}$$

If the density function belongs to the exponential family, the above properties lead to the following results

$$\mathbb{E}\left(\frac{\partial l(y,\theta)}{\partial\theta}\right) = \frac{1}{a(\phi)}\,\mathbb{E}\left(y - b'(\theta)\right) = 0,$$

and therefore

$$\mathbb{E}\left(y\right) = b'(\theta) = \mu. \tag{3.3}$$

Furthermore, the second property gives

$$\mathbb{E}\left(\frac{\partial^2 l(y,\theta)}{\partial\theta^2}\right) + \mathbb{E}\left(\left(\frac{\partial l(y,\theta)}{\partial\theta}\right)^2\right) = -\frac{1}{a(\phi)}b''(\theta) + \frac{1}{a(\phi)^2}\operatorname{var}(y) = 0,$$

and as a consequence we get

$$\operatorname{var}(y) = a(\phi)b''(\theta) = a(\phi)V(\mu), \tag{3.4}$$

where $V(\mu) = \frac{\partial\mu}{\partial\theta} = b''(\theta)$ is the so called variance function and $a(\phi)$ describes the dispersion.

## 3.2. Model Estimation

The aim of this section is to derive the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ of the model parameter $\boldsymbol{\beta}$. Thereafter, some properties of the estimate will be analysed.

From now on we assume that the responses $y_i$ are independent and from the same member of the exponential family with parameter $\theta_i$. In addition, the resulting expression for the variance, see (3.4), is now simplified to

$$\operatorname{var}(y_i) = \phi V(\mu_i),$$

where $a(\phi) = \phi$. Since in all the above examples of the exponential family this assumption holds, it constitutes no restriction for them.

The sample log-likelihood is therefore given by

$$l(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left( \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right).$$

To get the score function, we first take a look at

$$\frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial \mu_i} = \frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\phi} \left( y_i - b'(\theta_i) \right) \frac{\partial \theta_i}{\partial \mu_i}.$$

Since

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = V(\mu_i),$$

it follows that

$$\frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial \mu_i} = \frac{1}{\phi} \frac{y_i - \mu_i}{V(\mu_i)}, \tag{3.5}$$

where $b'(\theta_i) = \mu_i$, because of (3.3).

Since $\mu_i = \mu_i(\boldsymbol{\beta})$, we need to calculate

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \boldsymbol{x}_i = \boldsymbol{x}_i \frac{1}{g'(\mu_i)},$$

and get the desired result. Using the chain rule and the result in (3.5), the score function takes the form

$$\frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} = 0, \; j = 0, \ldots, p-1, \tag{3.6}$$

where the sum over $i = 1, \ldots, n$ is due to the fact that $\mu_i = \mu_i(\boldsymbol{\beta})$.

Since (3.6) can not be solved analytically, we settle for an iterative approach. In this case we use the Newton Raphson method and get

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left( -\frac{\partial^2 l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}^T} \right)^{-1} \frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}},$$

where all terms on the right-hand side are evaluated in $\boldsymbol{\beta}^{(t)}$, the result of the $t$-th iteration.

As one can see, the negative second derivative of the log-likelihood sample is necessary

for this iterative method. In our case it is given by

$$\begin{aligned}
-\frac{\partial^2 l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\beta_j \partial\beta_k} &= -\sum_{i=1}^{n} \frac{-\frac{x_{ik}}{g'(\mu_i)} V(\mu_i) - (y_i - \mu_i)\, V'(\mu_i) \frac{x_{ik}}{g'(\mu_i)}}{\phi V(\mu_i)^2} \frac{x_{ij}}{g'(\mu_i)} \\
&\quad - \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{-x_{ij}}{g'(\mu_i)^2} g''(\mu_i) \frac{x_{ik}}{g'(\mu_i)} \\
&= \sum_{i=1}^{n} \left( \frac{x_{ij}}{\phi V(\mu_i) g'(\mu_i)^2} + \frac{x_{ij}\,(y_i - \mu_i)\, V'(\mu_i)}{\phi V(\mu_i)^2 g'(\mu_i)^2} \right) x_{ik} \\
&\quad + \sum_{i=1}^{n} \frac{x_{ij}\,(y_i - \mu_i)\, g''(\mu_i)}{\phi V(\mu_i) g'(\mu_i)^3} x_{ik},
\end{aligned}$$

where

$$\frac{\partial V(\mu_i)}{\partial \eta_i} = \frac{\partial V(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} = V'(\mu_i) \frac{1}{g'(\mu_i)}.$$

If we define

$$\begin{aligned}
\frac{1}{w_i} &= V(\mu_i) g'(\mu_i)^2, \\
d_i &= g'(\mu_i),
\end{aligned}$$

then the negative Hessian matrix can be written as

$$-\frac{\partial^2 l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\beta_j \partial\beta_k} = \frac{1}{\phi} \sum_{i=1}^{n} x_{ij} \left( w_i + (y_i - \mu_i)\, w_i \left( \frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)} \right) \right) x_{ik}. \tag{3.7}$$

By defining $\boldsymbol{D}$ and $\boldsymbol{W}$ as diagonal matrices with entries $d_i$ and $w_i$, respectively, the score function in (3.6) can be written as

$$\frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\, \boldsymbol{\beta}} = \frac{1}{\phi}\, \boldsymbol{X}^T\, \boldsymbol{D}\, \boldsymbol{W}\, (\boldsymbol{y} - \boldsymbol{\mu}).$$

Next we define $\boldsymbol{W}^*$ as diagonal matrix with elements

$$w_i^* = w_i + (y_i - \mu_i)\, w_i \left( \frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)} \right).$$

It is easy to see that $\mathbb{E}\,(\boldsymbol{W}^*) = \boldsymbol{W}$. As a consequence, the negative second derivative in (3.7) takes the form

$$-\frac{\partial^2 l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\, \boldsymbol{\beta}\, \partial\, \boldsymbol{\beta}^T} = \frac{1}{\phi}\, \boldsymbol{X}^T\, \boldsymbol{W}^*\, \boldsymbol{X},$$

and the iterative method summarizes to

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left( \boldsymbol{X}^T\, \boldsymbol{W}^*\, \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T\, \boldsymbol{D}\, \boldsymbol{W}\, (\boldsymbol{y} - \boldsymbol{\mu}).$$

Including some pseudo-observations

$$\boldsymbol{z} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{W}^{*\,-1}\,\boldsymbol{D}\,\boldsymbol{W}\,(\boldsymbol{y} - \boldsymbol{\mu})\,, \tag{3.8}$$

the iteration changes to

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}^T\,\boldsymbol{W}^*\,\boldsymbol{X}\right)^{-1}\,\boldsymbol{X}^T\,\boldsymbol{W}^*\,\boldsymbol{z}\,. \tag{3.9}$$

This method is called iterative weighed least squares or IWLS, since in each iteration a weighed least squares problem needs to be solved. In addition, one might notice that in the final version of the iteration $\phi$ is cancelled, indicating that $\phi$ is of no importance regarding the estimation of $\boldsymbol{\beta}$. This conclusion is supported by (3.6), where $\phi$ can already be neglected.

**Definition.** *Fisher Scoring Technique*
*It is customary to use* $\mathbb{E}\,(\boldsymbol{W}^*) = \boldsymbol{W}$ *instead of* $\boldsymbol{W}^*$ *in* (3.9) *and thereby get the simpler iteration relation*

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

*where*

$$\boldsymbol{z} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{D}\,(\boldsymbol{y} - \boldsymbol{\mu})\,.$$

*This approach is called the Fisher scoring technique.*

A justification of the Fisher scoring technique will be given now. To do so, we introduce an alternative way to find the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, and in the end we show that this method is equivalent to the Fisher scoring technique.

The now presented alternative way to get the estimator $\hat{\boldsymbol{\beta}}$ can be found in Wood (2006a). We introduce it at this point because it will be used later to estimate the parameters of a generalized additive model. Earlier, we defined $\hat{\boldsymbol{\beta}}$ as the solution of the estimating equation

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i - \mu_i)}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j} = 0, \tag{3.10}$$

or see (3.6).

Looking at the first derivative of

$$\mathcal{S}_g = \frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i - \mu_i(\boldsymbol{\beta}))^2}{V(\mu_i)},$$

namely

$$\frac{\partial\mathcal{S}_g}{\partial\beta_j} = \frac{2}{\phi}\sum_{i=1}^{n}\frac{(y_i - \mu_i(\boldsymbol{\beta}))}{V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j},$$

where $V(\mu_i)$ is treated as fixed, one notices that setting the first derivative of $\mathcal{S}_g$ to zero, and therefore minimizing $\mathcal{S}_g$, is equivalent to solving (3.10). In addition, we observe that $\phi$ is not important to the estimation process and can therefore be disregarded.

If we define $\boldsymbol{V}$ as diagonal matrix with entries $v_{ii} = V(\mu_i)$ and neglect $\phi$, then $\mathcal{S}_g$ can be written as

$$\mathcal{S}_g = \| \boldsymbol{V}^{-1/2}(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))\|^2.$$

The minimization of $\mathcal{S}_g$ is achieved by exerting the methods of Chapter 2 in an iterative way. Starting with $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\eta}^{(0)} = \boldsymbol{X}\,\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\mu}^{(0)} = g^{-1}(\boldsymbol{\eta}^{(0)})$, where $g^{-1}(\cdot)$ describes the inverse link function, the iteration method is summarized in the following three steps:

- $\boldsymbol{\eta}^{(k)}$ and $\boldsymbol{\mu}^{(k)}$ are calculated by means of $\boldsymbol{\beta}^{(k)}$. Thereafter, $\boldsymbol{V}^{(k)} = V(\boldsymbol{\mu}^{(k)})$ can also be computed.
- $\mathcal{S}_g$ can be minimized like the sum of squares of a linear model. As a result of the minimization we obtain $\boldsymbol{\beta}^{(k+1)}$.
- $k \to k+1$

According to Wood (2006a) this method sometimes poses a problem because $\boldsymbol{\beta}^{(k)}$ can converge a lot faster than $V(\boldsymbol{\mu}^{(k)})$. This makes sense, if we consider that $V(\boldsymbol{\mu}^{(k)})$ is always calculated with $\boldsymbol{\beta}^{(k)}$, while in the same iteration $\boldsymbol{\beta}^{(k+1)}$ is computed. As a result, $V(\boldsymbol{\mu}^{(k)})$ is always one step behind. Therefore, we need to slow down the process for $\boldsymbol{\beta}^{(k)}$ by introducing a Taylor expansion of $\boldsymbol{\mu}(\boldsymbol{\beta})$ in $\boldsymbol{\beta}^{(k)}$. As a result, $\mathcal{S}_g$ is approximately

$$\begin{aligned}
\mathcal{S}_g &\approx \left\| \left(\boldsymbol{V}^{(k)}\right)^{-1/2} \left( \boldsymbol{y} - \boldsymbol{\mu}^{(k)} - \frac{\partial\,\boldsymbol{\mu}}{\partial\,\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}^{(k)}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\right) \right) \right\|^2 \\
&\approx \left\| \left(\boldsymbol{V}^{(k)}\right)^{-1/2} \left( \boldsymbol{y} - \boldsymbol{\mu}^{(k)} - \left(\boldsymbol{D}^{(k)}\right)^{-1} \boldsymbol{X} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}\right) \right) \right\|^2,
\end{aligned}$$

where $\boldsymbol{D}^{(k)}$ is a diagonal matrix with elements $d_{ii} = g'(\mu_i^{(k)})$ and $\boldsymbol{\beta}$ denotes the true parameter. Next, we can summarize the approximation of $\mathcal{S}_g$ as follows

$$\begin{aligned}
\mathcal{S}_g &\approx \| \left(\boldsymbol{V}^{(k)}\right)^{-1/2} \left(\boldsymbol{D}^{(k)}\right)^{-1} \left( \boldsymbol{D}^{(k)} \left( \boldsymbol{y} - \boldsymbol{\mu}^{(k)} \right) + \boldsymbol{X}\,\boldsymbol{\beta}^{(k)} - \boldsymbol{X}\,\boldsymbol{\beta} \right) \|^2 \\
&\approx \| \left(\boldsymbol{W}^{(k)}\right)^{1/2} \left( \boldsymbol{z}^{(k)} - \boldsymbol{X}\,\boldsymbol{\beta} \right) \|^2, \qquad (3.11)
\end{aligned}$$

where $\boldsymbol{W}^{(k)}$ and $\boldsymbol{z}^{(k)}$ are defined as

$$\begin{aligned}
w_{ii}^{(k)} &= \frac{1}{V(\mu_i^{(k)}) g'(\mu_i^{(k)})^2}, \\
z_i^{(k)} &= g'(\mu_i^{(k)}) \left( y_i - \mu_i^{(k)} \right) + \boldsymbol{x}_i^T\,\boldsymbol{\beta}^{(k)}.
\end{aligned}$$

Taking the approximation above into account, the iteration method is now given by:

- Given a current $\boldsymbol{\beta}^{(k)}$, we calculate $\boldsymbol{\eta}^{(k)}$, $\boldsymbol{\mu}^{(k)}$ and $\boldsymbol{V}^{(k)}$.
- Minimize (3.11) like a linear model and obtain $\boldsymbol{\beta}^{(k+1)}$.
- $k \to k+1$

Now we take a closer look at the second step of this iteration method. If we assume that $\boldsymbol{W}^{(k)}$ and $\boldsymbol{z}^{(k)}$ in (3.11) are independent of the true parameter $\boldsymbol{\beta}$, then the first derivative of $\mathcal{S}_g$ with respect to $\boldsymbol{\beta}$ takes the form

$$\frac{\partial \mathcal{S}_g}{\partial \boldsymbol{\beta}} \approx 2\, \boldsymbol{X}^T \boldsymbol{W}^{(k)} \left(\boldsymbol{z}^{(k)} - \boldsymbol{X}\, \boldsymbol{\beta}\right) = \boldsymbol{0}\,.$$

By setting this first derivative to zero, we get the following equation as solution of the minimization problem

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

Therefore, the estimate $\boldsymbol{\beta}^{(k+1)}$ is again given by

$$\boldsymbol{\beta}^{(k+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}^{(k)}\, \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(k)}\, z^{(k)},$$

where $\boldsymbol{W}^{(k)}$ and $\boldsymbol{z}^{(k)}$ are defined as above.

The derivation of the estimate above shows that this iteration method is the same as the Fisher scoring technique. Thereby, this method is also similar to the iterative method introduced previously, if we replace $\boldsymbol{W}^*$ by its mean $\mathbb{E}\left[\boldsymbol{W}^*\right] = \boldsymbol{W}$. This alternative way to get the estimator $\hat{\boldsymbol{\beta}}$ also justifies the use of the Fisher Scoring Technique and allows for a better understanding what changes if we replace $\boldsymbol{W}^*$ by $\boldsymbol{W}$. It is the difference between minimizing $\mathcal{S}_g$ or minimizing $\mathcal{S}_g$ in respect to a Taylor expansion of $\boldsymbol{\mu}(\boldsymbol{\beta})$.

## 3.3. Asymptotic Behaviour

As in Chapter 2 we are interested in some distributional results of the estimator $\hat{\boldsymbol{\beta}}$. For linear models it was possible to derive those results explicitly, while in the case of generalized linear models we get them only asymptotically by using the Taylor expansion in the true parameter value $\boldsymbol{\beta}$, i.e.

$$\boldsymbol{0} = \left.\frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\, \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} \approx \left.\frac{\partial l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\, \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}} + \left.\frac{\partial^2 l(\boldsymbol{y}, \boldsymbol{\theta}(\boldsymbol{\beta}))}{\partial\, \boldsymbol{\beta}\, \partial\, \boldsymbol{\beta}^T}\right|_{\boldsymbol{\beta}} \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

It follows that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \left(\boldsymbol{X}^T \boldsymbol{W}\, \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{D}\, \boldsymbol{W}\, (\boldsymbol{y} - \boldsymbol{\mu})\,,$$

where $\boldsymbol{W}$, $\boldsymbol{D}$ and $\boldsymbol{\mu}$ are evaluated in the true parameter $\boldsymbol{\beta}$. As a result, the mean and the variance of $\hat{\boldsymbol{\beta}}$ are approximately

$$\begin{aligned}\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &\approx \boldsymbol{\beta},\\ \operatorname{var}(\hat{\boldsymbol{\beta}}) &\approx \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{W}\operatorname{var}(\boldsymbol{y})\boldsymbol{W}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\\ &\approx \phi\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1},\end{aligned}$$

since $\operatorname{var}(\boldsymbol{y}) = \phi\left(\boldsymbol{D}\boldsymbol{W}\boldsymbol{D}\right)^{-1}$. Fahrmeir and Kaufmann (1985) did even show that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \overset{n\to\infty}{\sim} \mathcal{N}\left(\mathbf{0}, n\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\right).$$

## 3.4. Pearson Statistic

Since in the last section $\boldsymbol{W}$ was evaluated in the true parameter $\boldsymbol{\beta}$, we deduce an estimate for $\operatorname{var}(\hat{\boldsymbol{\beta}})$ by using the estimate $\hat{\boldsymbol{\beta}}$ instead, i.e.

$$\widehat{\operatorname{var}}\left(\hat{\boldsymbol{\beta}}\right) = \phi\left(\boldsymbol{X}^T\boldsymbol{W}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\right)^{-1}.$$

But still in this expression the parameter $\phi$ is not always known. Therefore, an estimator for $\phi$ needs to be established. We already know that the dispersion parameter can be written as $\phi = \operatorname{var}(y_i)/V(\mu_i)$ and therefore if $\boldsymbol{\beta}$ is known we get the unbiased estimator

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

For $\boldsymbol{\beta}$ unknown the bias corrected form of this estimator is

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

where $X^2$ is called the Pearson statistic.

In the same way the Pearson residuals are given by

$$r_i = \frac{y_i-\hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}.$$

The Pearson residuals are one possibility to standardize the residuals. According to Wood (2006a) they should have approximately zero mean and variance $\phi$, if the model is correct.

## 3.5. Deviance

The scaled deviance is a measurement of goodness of fit and defined by

$$\frac{1}{\phi} D\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}\right) = -2\left(l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) - l(\boldsymbol{y}, \boldsymbol{y})\right). \tag{3.12}$$

One may observe that the deviance is the difference between the log-likelihood of the fitted model and the log-likelihood of a saturated model. Since in a saturated model there are as many parameters included as observations, namely $n$, the fitted values $\hat{\mu}_i$ are equal to $y_i$, $i = 1, \ldots, n$. Furthermore, because of the independence of $l(\boldsymbol{y}, \boldsymbol{y})$ in (3.12) from $\hat{\boldsymbol{\mu}}$ and thereby from the estimated model, $l(\boldsymbol{y}, \boldsymbol{y})$ is constant if $\boldsymbol{y}$ is given. As a result, the maximum likelihood estimate $\hat{\boldsymbol{\mu}}$ maximizes the likelihood function $l(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ and, thus, minimizes the deviance.

On the other hand, the deviance is some sort of a generalisation of the residual sum of squares in a linear model. If we assume normally distributed responses as in a linear regression model, $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, then the log-likelihood is

$$l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)^2.$$

As a result, the scaled deviance takes the form

$$\begin{aligned}\frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})) &= -2\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)^2 \right.\\ &\qquad \left. + \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - y_i\right)^2\right)\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)^2,\end{aligned}$$

where $\phi \equiv \sigma^2$ and the last result is equivalent to the residual sum of squares.

In addition, from the definition of the scaled deviance and the approximate $\chi^2_{n-p}$ distribution thereof, see Mc Cullagh and Nelder (1989), another unbiased estimator for $\phi$ follows, i.e.

$$\hat{\phi} = \frac{1}{n-p} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}).$$

Since the mean of $\frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is $n - p$, the unbiased estimator above results. For normally distributed values the $\chi^2$ distribution is true, but for other distributions this only holds approximately.

## 3.6. Hypothesis Testing

As in the case of linear models we want to test if all $p$ explanatory variables are necessary in the predictor or if a subset $p-q$ suffices. Therefore, we test the following hypothesis

$$\begin{aligned} H_0 &: \beta_1 = \cdots = \beta_q = 0, \\ H_1 &: \beta_1, \ldots, \beta_q \neq 0. \end{aligned}$$

Let $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ denote the fitted values under both these hypotheses, respectively. Then the likelihood ratio test statistic is given by

$$\begin{aligned} -2\left(l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)\right) &= -2\Big(l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - l(\boldsymbol{y}, \boldsymbol{y}) - l(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1) + l(\boldsymbol{y}, \boldsymbol{y})\Big) \\ &= \frac{1}{\phi}\Big(D\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1\right)\Big). \end{aligned}$$

If $\phi$ is known than we can test the hypothesis by calculating

$$\frac{1}{\phi}\left(D\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1\right)\right) \sim \chi^2_q,$$

which only holds approximately and in the large sample limit. Mc Cullagh and Nelder (1989) state that in this case the $\chi^2$ distribution holds for not normally distributed responses better than in the last section above. Otherwise we use the statistic

$$\frac{\left(D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)\right)/q}{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)/\left(n-p\right)} \sim F_{q,n-p},$$

i.e. we replace $\phi$ by its estimator $\frac{1}{n-p}D\left(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1\right)$, which then tends to a statistic that might be approximated by a $F$ distribution.

# 4. Additive Models

In the same way as linear models have their restrictions, generalized linear models (GLMs) sometimes lack some necessary flexibility to adequately describe the structure of the data. Before we introduce generalized additive models (GAMs) in the next chapter, we want to talk about additive models. This model class is similar to linear regression models, in fact we will show that certain additive models can be reduced to linear models. But contrary to the linear models in Chapter 2 the mean of the response $y$ is now not a linear function of a continuous explanatory variable $x$ but nonlinear in $x$, i.e.

$$y = f(x) + \varepsilon, \tag{4.1}$$

where $(x, y)$ represents one observation of a dataset, $y \sim \mathcal{N}(\mu, \sigma^2)$ and $f(\cdot)$ is a smooth function. A function is called smooth if the requirements of continuity and of continuous first and second derivatives are fulfilled.

During this chapter we will introduce different possibilities of how to estimate $f(\cdot)$. But all possibilities have in common that we demand that $f(\cdot)$ can be written as

$$f(x) = \sum_{j=0}^{q-1} \gamma_j b_j(x), \tag{4.2}$$

where $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{q-1})^T$ describes the parameters and $b_j(x)$ stands for a basis function. In a function space basis functions build a basis similar to vectors in a vector space by linear independence and the spanning property. Therefore, we search functions that form a basis in our function space by fulfilling these properties. Since we want to estimate the smooth function $f(\cdot)$ by polynomial functions, our function space is the polynomial function space. In the following, we will introduce different bases and will discuss their various properties.

As a consequence of (4.2) model (4.1) is linear in the parameters and is therefore equivalent to a linear regression model. Consequently model estimation is similar as in Chapter 2, but before we illustrate the issue by various examples of basis functions, we want to introduce the data which is used in the following.

The dataset is obtained from the Open Grid Europe GmbH (OGE), a leading German gas transmission operator. For more details about this project we refer to Friedl et al. (2012). The dataset contains the information of one knot of the German gas network and combines several variables, some of which are

- `temp` ... the daily mean temperature in degree Celsius, and
- `max.flow` ... the daily maximal gas flow in KWh per hour.

Since the German gas transmission operators have to guarantee sufficient gas flow as low as $-12$ degree Celsius, they are interested in the mean behaviour the maximum gas flow takes with decreasing temperature. As daily mean temperatures as lows as $-12°C$ have not been observed in the past, one part of Chapter 5 will concentrate on the theory how to extrapolate a given model. The application of the theory on the data can be found in Chapter 6.

To get a first impression of the structure of the data the analysed data is shown in Figure 4.1. We can observe that with increasing temperature the maximum gas flow continuously decreases and that above 15°C the decrease lessens until it is almost constant. Next we will try to estimate this relationship through additive models using different basis functions.

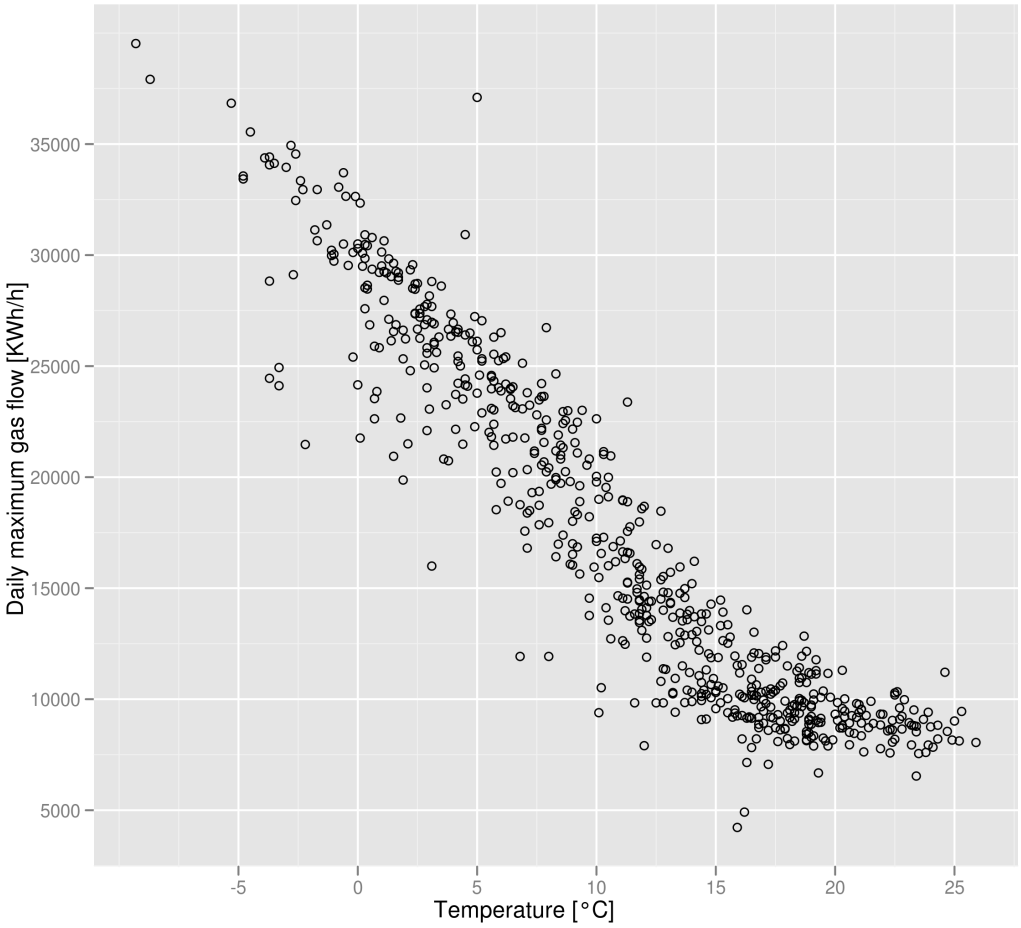


Figure 4.1.: Daily maximum gas flow plotted against the daily mean temperature.

# 4.1. Basis Functions

It is the aim of this section to present possible choices of basis functions $b_j(x)$ for the function in (4.2) and to discuss their various advantages and disadvantages. We start by estimating a polynomial over the whole data range and continue with splines. For every choice of basis functions besides their definition the R-code to calculate them and to estimate the fit is given. In addition, plots of the basis functions and the resulting fits are added.

## 4.1.1. Polynomial Basis

The initial choice for the basis functions $b_j(x)$ in (4.2) is a set of $q$ polynomials. In that sense the first basis we want to discuss is

$$\begin{aligned} &b_0(x) = 1, && b_1(x) = x, && b_2(x) = x^2, \\ &b_3(x) = x^3, && b_4(x) = x^4, && \dots, \end{aligned}$$

and therefore

$$f(x) = \sum_{j=0}^{q-1} \gamma_j x^j. \tag{4.3}$$

The first four of these basis functions can be observed in Figure 4.2 for $x \in [0, 1]$.

To get the smooth function $f(\cdot)$ resulting from this choice of basis functions, the parameters $\gamma_j$ in (4.3) need to be estimated. In addition, the maximal number of polynomials $q$, or the maximal polynomial degree $m = q - 1$ needs to be chosen. For $q$ fixed the parameters $\gamma_j$ can be easily estimated using the statistical software R (see R-code below).

If we assume normally distributed responses and a model as in (4.1), where $f(\cdot)$ is defined as in (4.3), then $f(\cdot)$ is linear in the parameters $\gamma_j$. Therefore, we are able to reduce model (4.1) to a linear regression model. As a consequence, the theory of Chapter 2 applies and the parameters $\gamma_j$ can be estimated through

```
q<-4
X1<-outer(temp,0:(q-1),"^") #model matrix
mod.poly1<-lm(max.flow~X1-1) #fit model

p<-ggplot(pday_small,aes(temp,max.flow)) +geom_point(shape=1) #plot
p+geom_line(aes(x=temp,y=X1%*%coef(mod.poly1)),colour="red") # +fit
```

In the above R-code the command `lm` estimates the parameters of a linear model with model matrix `X1`. Since the intercept is already included in `X1` as the polynomial of degree zero, it does not need to be added, hence `-1`. The last two lines of the R-code above generate the plots shown in Figure 4.3. For the remaining part of this chapter in the R-code the plot `p` describes the basis plot including only the data points, while the

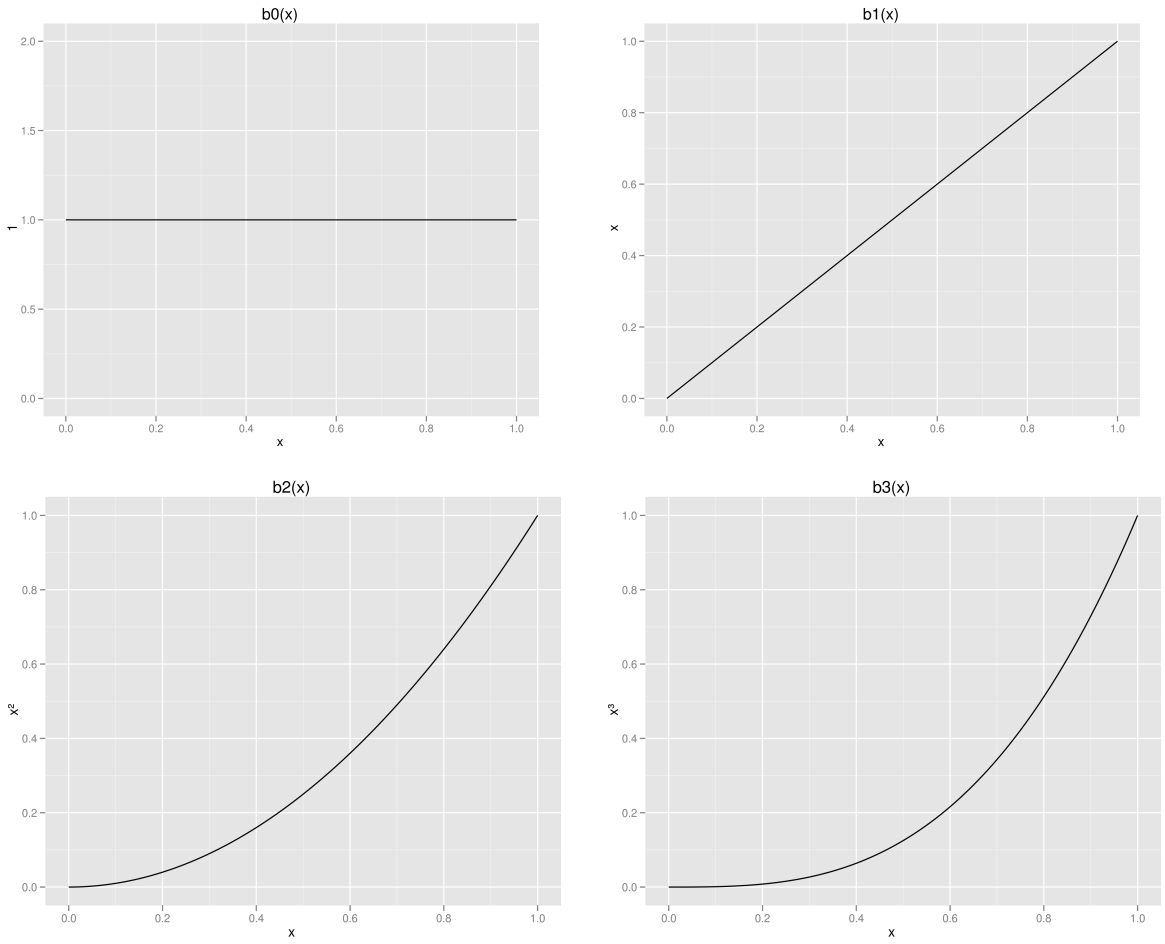


Figure 4.2.: Plot of the first four polynomial basis functions for $x \in [0, 1]$.

fits are usually added using the command `geom_line`. Since all plots in this work have been produced by the R package `ggplot2`, it needs to be loaded before these plots can be generated. More information about the package `ggplot2` can be found in Wickham (2009).

One advantage of this approach is that it is very simple. The task to estimate for example the parameters of such a cubic function of the form

$$f(x) = \gamma_1 + \gamma_2 x + \gamma_3 x^2 + \gamma_4 x^3$$

is not really challenging for most people with a mathematical background and the knowledge of a statistical or mathematical software. Another advantage to other choices of basis functions is that it is possible to explicitly write down in full the resulting function. Although this approach performs well for polynomial functions, in general it sometimes lacks the necessary flexibility to describe the mean of a response variable.

If we apply this R-code to the gas flow dataset assuming $m = q - 1 = 3$, then the upper left plot in Figure 4.3 results. Especially in the upper left corner of this plot it seems as if the model is not flexible enough to describe the data properly.

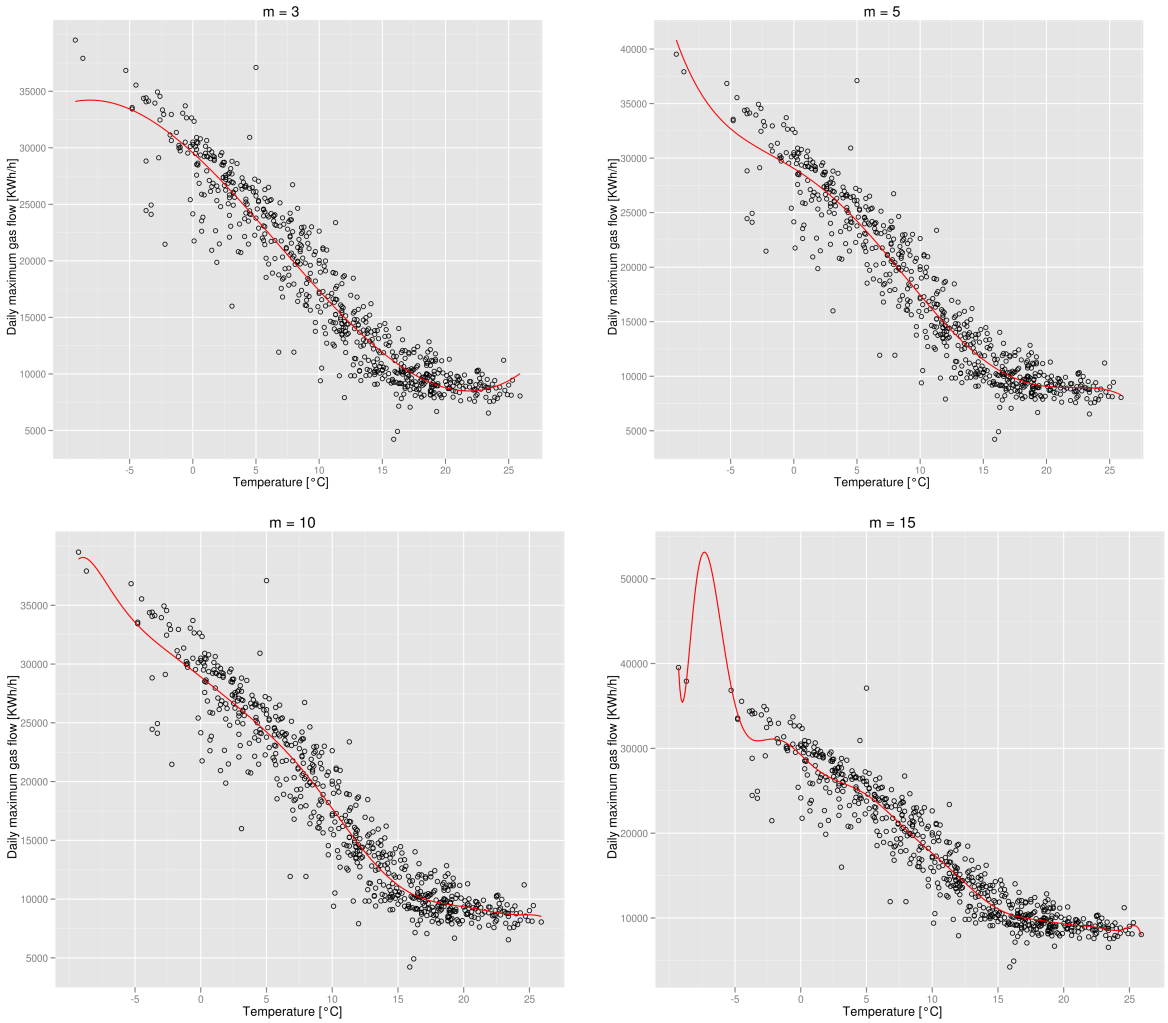


Figure 4.3.: Smooth fits resulting from polynomial basis functions with degree $m = 3, 5, 10, 15$.

Since a higher polynomial degree adds flexibility to the model, we estimate the same model for $m = 5$ and $m = 10$. The results are displayed in Figure 4.3 in the upper right and lower left plot. In the lower right plot in Figure 4.3 a polynomial with degree $m = 15$ is shown. One might notice that in the upper left corner where few observations are found the fit is very wiggly. All in all, it seems as if this approach is not able to catch the structure of the data in a satisfying way.

### 4.1.2. Splines

Splines are the logical extension of polynomial basis functions. While in the last section one polynomial was fitted over the whole data range, now the idea is to gain flexibility by using polynomials of degree $m$ in subintervals of the explanatory variable $x$. Therefore,

a set of knots $(\kappa_1, \ldots, \kappa_K)$ is chosen, which partitions the range $[a, b]$ of the covariate into intervals $[a, \kappa_1], [\kappa_1, \kappa_2], \ldots, [\kappa_K, b]$, where $a$ and $b$ stand for the left and right end of the $x$ range, respectively. For the $j$-th interval a polynomial of degree $m$, further referred to as $f_{[j]}(x)$ can now be estimated. To ensure a certain degree of continuity and differentiability, continuous derivatives up to order $m-1$ are required at all points of adjoining intervals.

For example if $m = 3$ and $f_{[j]}(x) = f(x)|_{\kappa_{j-1}}^{\kappa_j}$, therefore describing a cubic polynomial on the interval $[\kappa_{j-1}, \kappa_j]$, then we demand that

$$\begin{aligned} f_{[j]}(\kappa_j) &= f_{[j+1]}(\kappa_j), \quad j = 1, \ldots, K, \\ f'_{[j]}(\kappa_j) &= f'_{[j+1]}(\kappa_j), \quad j = 1, \ldots, K, \\ f''_{[j]}(\kappa_j) &= f''_{[j+1]}(\kappa_j), \quad j = 1, \ldots, K. \end{aligned}$$

As a result, we gain a smooth function $f(x)$.

In the following, three different possibilities for spline basis functions will be discussed.

### Truncated Power Series Basis

The first approach, which is also closest to the polynomial approach of the last subsection, are truncated power (TP) series.

For a given set of knots $(\kappa_1, \ldots, \kappa_K)$ a TP-series of degree $m$ has the form

$$f(x) = \sum_{k=0}^{m} \gamma_k x^k + \sum_{k=1}^{K} \gamma_{k+m} (x - \kappa_k)_+^m, \tag{4.4}$$

where

$$(x - \kappa_k)_+ = \begin{cases} x - \kappa_k & \text{if } x > \kappa_k, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from (4.4) that each interval $[\kappa_k, \kappa_{k+1}]$ has its own polynomial of degree $m$. For example on the first two intervals $f(\cdot)$ is given by

$$\begin{aligned} f_{[1]}(x) &= \sum_{k=0}^{m} \gamma_k x^k, && x \in [a, \kappa_1], \\ f_{[2]}(x) &= \sum_{k=0}^{m} \gamma_k x^k + \gamma_{1+m} (x - \kappa_1)^m, && x \in [\kappa_1, \kappa_2]. \end{aligned}$$

If we evaluate $f_{[j]}(x)$ and $f_{[j+1]}(x)$ in $\kappa_j$, then

$$\begin{aligned} f_{[j]}(\kappa_j) &= \sum_{k=0}^{m} \gamma_k (\kappa_j)^k + \sum_{k=1}^{j-1} \gamma_{k+m} (\kappa_j - \kappa_k)^m, \\ f_{[j+1]}(\kappa_j) &= \sum_{k=0}^{m} \gamma_k (\kappa_j)^k + \sum_{k=1}^{j} \gamma_{k+m} (\kappa_j - \kappa_k)^m, \end{aligned}$$

and the last term of the second sum of $f_{[j+1]}(\kappa_j)$, namely $\gamma_{j+m}\,(\kappa_j - \kappa_j)^m$, is zero. Therefore, $f_{[j]}(\kappa_j) = f_{[j+1]}(\kappa_j), j = 1, \ldots, K$, and the continuity of $f(x)$ over the entire $x$ range follows. In the same way, the continuity of the first $m-1$ derivatives ensues, while the $m$-th derivative takes the form of a step function. As a result, the above requirements for splines are satisfied.

The first $q\ (= m+1)$ basis functions of the TP-series are the same as in the polynomial approach, while the last $K$ basis functions are added in this method. Therefore, in Figure 4.4 only the five new basis functions resulting from $m = 3$ and $K = 5$ are shown. While the black lines in Figure 4.4 represent the basis functions, the dashed grey lines mark the position of the knots and $a$ and $b$, respectively.

If we assume the same model as in the polynomial case (see (4.1)) with normally distributed responses, then the following R-code estimates the parameters $\boldsymbol{\gamma}$.

```
m<-3; K<-5

Xtp1<-outer(temp,0:m,"^") #first part of model matrix

knots<-seq(min(temp),max(temp),length.out=(K+2))

Xtp2<-matrix(NA,length(temp),K) #second part of model matrix
for(k in 2:(K+1)) { #K+1 = number of intervals
    knot_k<-knots[k]
    Xtp2[,k-1]<-(pmax((temp-knot_k),rep(0,length(temp))))^m
}

Xtp<-cbind(Xtp1,Xtp2) #model matrix

mod.tp1<-lm(max.flow ~ Xtp -1) #fit model

line_knots<-geom_vline(xintercept=knots,linetype=2,alpha=1/3)
p + line_knots                    # plot data, knots and fit
  + geom_line(aes(x=temp,y=Xtp%*%coef(mod.tp1)),colour="red")
```

In this R-code the first half of the model matrix (`Xtp1`) is defined in the same way as the model matrix in the polynomial case. Thereafter, the knots including $a$ and $b$ are set equidistantly and the remaining basis functions $(x-\kappa_k)_+^m$ are calculated and summarized in `Xtp2`. The model matrix of the TP-series results from the combination of `Xtp1` and `Xtp2` in `Xtp`. Finally, we are able to estimate the parameters $\boldsymbol{\gamma}$ by using the R-function `lm`. At last the R-code for the plots of the fitted values is added. While `line_knots` describes the vertical lines for the knots, `p` is defined as before and `geom_line` draws the red line for the fitted values.

The fits that result from applying this code to the gas flow dataset are shown in Figure 4.5 for $m = 3$ and $K = 1, 5, 10, 15$. The red lines represent the different fits, while the grey dashed lines again show the position of the knots.

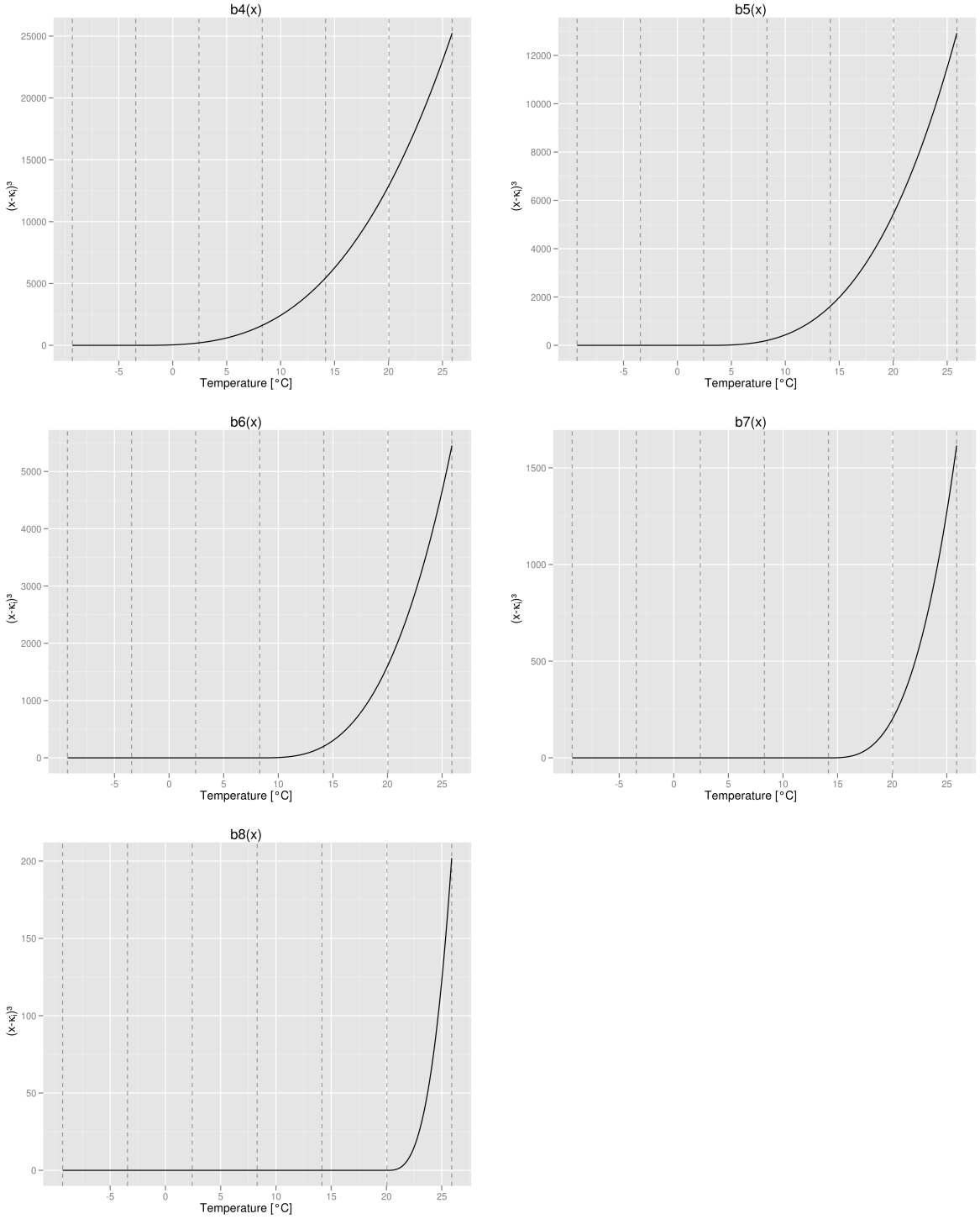


Figure 4.4.: Plot of the last five basis functions of a TP-series with $m = 3$ and $K = 5$.

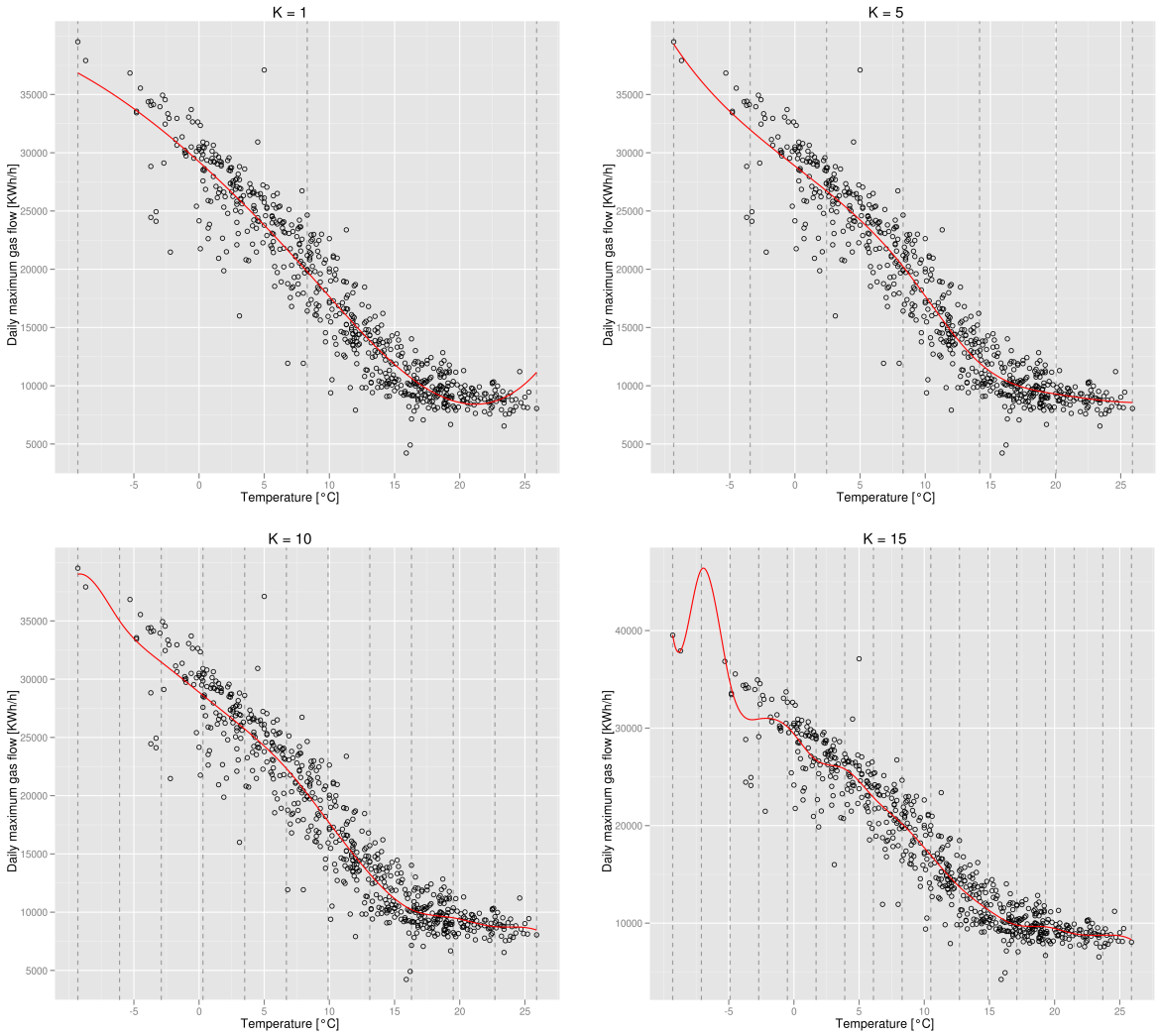


Figure 4.5.: Smooth fit resulting from TP-series basis functions with $m = 3$ and $K = 1, 5, 10, 15$.

We observe that while the fit in the upper left plot ($K = 1$) lacks the necessary flexibility to describe the data properly, the fits from the lower charts ($K = 10, 15$) are in our opinion too wiggly. Especially in the upper left corner of these plots where little data is found a higher number of knots leads to a fit that is far too fluctuating (see lower right plot). All in all the upper right plot ($K = 5$) seems to describe the data best.

This example demonstrates very well one of the disadvantages of this choice of basis functions. We have to choose the number of knots and as one could see in Figure 4.5 this choice greatly influences the fit.

On the other hand, the simplicity of the TP-series rates among their advantages and makes them easy to handle and easy to understand. Again we are also able to explicitly write down the resulting function $f(x)$ and its estimate, which is in many cases very useful since it makes interpretation a lot easier.

### Cubic Splines

A favoured choice of polynomials are cubic polynomials, and therefore cubic splines. Later in Theorem 1 a reason for this choice of polynomial degree is given. However, the aim of this part is to introduce another possible choice of basis functions for cubic splines Wood (2006a).

For any $x$ in the interval $[\kappa_j, \kappa_{j+1}]$, we define

$$\begin{aligned} f_{[j+1]}(x) &= \frac{\kappa_{j+1}-x}{h_j}\gamma_j + \frac{x-\kappa_j}{h_j}\gamma_{j+1} + \frac{1}{6}\delta_j\left(\frac{(\kappa_{j+1}-x)^3}{h_j} - h_j(\kappa_{j+1}-x)\right) \\ &\quad + \frac{1}{6}\delta_{j+1}\left(\frac{(x-\kappa_j)^3}{h_j} - h_j(x-\kappa_j)\right), \quad j = 0, \ldots, K, \end{aligned} \tag{4.5}$$

with $\kappa_0 = a$ and $\kappa_{K+1} = b$ and where $h_j = \kappa_{j+1} - \kappa_j$, $j = 0, \ldots, K$. In (4.5) the parameters $\gamma_j$ and $\delta_j$ need to be estimated, but we will show that by satisfying the restrictions of a spline function $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_{K+1})^T$ can be expressed in terms of $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{K+1})^T$. Therefore, we will finally only need to estimate the parameters $\boldsymbol{\gamma}$. Hereafter, we show that $f(x)$ is continuous and has a continuous second derivative. In addition, from the definition of $f_{[j+1]}(x)$ in (4.5), its first and second derivative ensue

$$\begin{aligned} f'_{[j+1]}(x) &= \frac{1}{h_j}(\gamma_{j+1} - \gamma_j) + \frac{1}{6}\delta_j\left(-\frac{3}{h_j}(\kappa_{j+1}-x)^2 + h_j\right) \\ &\quad + \frac{1}{6}\delta_{j+1}\left(\frac{3}{h_j}(x-\kappa_j)^2 - h_j\right), \end{aligned} \tag{4.6}$$

$$f''_{[j+1]}(x) = \frac{\delta_j}{h_j}(\kappa_{j+1}-x) + \frac{\delta_{j+1}}{h_j}(x-\kappa_j), \tag{4.7}$$

with $x \in [\kappa_j, \kappa_{j+1}]$ and $j = 0, \ldots, K$.

By evaluating $f(x)$ at the inner knots $(\kappa_1, \ldots, \kappa_K)$, the continuity of $f$ follows. For example if we analyse $f_{[j+1]}(x)$ in $\kappa_j$, we get

$$f_{[j+1]}(\kappa_j) = \frac{\kappa_{j+1}-\kappa_j}{h_j}\gamma_j + \frac{1}{6}\delta_j\left(\frac{(\kappa_{j+1}-\kappa_j)^3}{h_j} - h_j(\kappa_{j+1}-\kappa_j)\right).$$

Since $h_j = \kappa_{j+1} - \kappa_j$, it follows that

$$f(\kappa_j) = f_{[j+1]}(\kappa_j) = \gamma_j + \frac{1}{6}\delta_j\left(h_j^2 - h_j^2\right) = \gamma_j,$$

where the first result is due to the definition of $f_{[j]}(x) = f(x)|_{\kappa_{j-1}}^{\kappa_j}$. If we evaluate $f_{[j]}(x)$ in $\kappa_j$, we get

$$\begin{aligned} f_{[j]}(\kappa_j) &= \frac{\kappa_j - \kappa_{j-1}}{h_{j-1}}\gamma_j + \frac{1}{6}\delta_j\left(\frac{(\kappa_j-\kappa_{j-1})^3}{h_{j-1}} - h_{j-1}(\kappa_j - \kappa_{j-1})\right) \\ &= \gamma_j + \frac{1}{6}\delta_j\left(h_{j-1}^2 - h_{j-1}^2\right) = \gamma_j = f(\kappa_j), \end{aligned}$$

and therefore $f_{[j+1]}(\kappa_j) = f_{[j]}(\kappa_j) = f(\kappa_j)$, $j = 1, \ldots, K$. We also found that $\gamma_j = f(\kappa_j)$, meaning that the $j$-th parameter can be interpreted as the function value at $\kappa_j$.

In the same way we can show that the second derivative $f''(x)$ is continuous. From the definition of the second derivative on the interval $[\kappa_j, \kappa_{j+1}]$ in (4.7), the continuity of the second derivative ensues if we evaluate $f''_{[j+1]}(x)$ and $f''_{[j]}(x)$ in $\kappa_j$, i.e.

$$f''(\kappa_j) = f''_{[j+1]}(\kappa_j) = \frac{\delta_j}{h_j}(\kappa_{j+1} - \kappa_j) = \delta_j,$$
$$f''(\kappa_j) = f''_{[j]}(\kappa_j) = \frac{\delta_j}{h_{j-1}}(\kappa_j - \kappa_{j-1}) = \delta_j,$$

where the first result is again due to the definition of $f_{[j]}(x)$, and therefore $f''_{[j+1]}(\kappa_j) = f''_{[j]}(\kappa_j) = f''(\kappa_j)$, $j = 1, \ldots, K$. Furthermore, $\delta_j = f''(\kappa_j)$ and can therefore be interpreted as the second derivative of $f(x)$ in $\kappa_j$.

Now the only missing restriction for a cubic spline is the continuity of the first derivative, which is defined in (4.6). By evaluating $f'_{[j]}(x)$ and $f'_{[j+1]}(x)$ in $\kappa_j$, we get

$$f'_{[j+1]}(\kappa_j) = \frac{1}{h_j}(\gamma_{j+1} - \gamma_j) - \frac{1}{3}h_j\delta_j - \frac{1}{6}h_j\delta_{j+1},$$
$$f'_{[j]}(\kappa_j) = \frac{1}{h_{j-1}}(\gamma_j - \gamma_{j-1}) + \frac{1}{6}h_{j-1}\delta_{j-1} + \frac{1}{3}h_{j-1}\delta_j.$$

From the claim of $f'_{[j+1]}(\kappa_j) = f'_{[j]}(\kappa_j)$, it follows that

$$\frac{1}{h_j}\gamma_{j+1} + \left(-\frac{1}{h_j} - \frac{1}{h_{j-1}}\right)\gamma_j + \frac{1}{h_{j-1}}\gamma_{j-1} = \frac{1}{6}h_j\delta_{j+1} + \frac{1}{3}(h_j + h_{j-1})\delta_j + \frac{1}{6}h_{j-1}\delta_{j-1},$$

$j = 1, \ldots, K$. The restriction above can also be written in matrix representation as

$$\boldsymbol{G}\,\boldsymbol{\gamma} = \boldsymbol{B}\,\boldsymbol{\delta}^*, \tag{4.8}$$

with $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{K+1})^T$ and $\boldsymbol{\delta} = (\delta_0, \ldots, \delta_{K+1})^T = (0, \boldsymbol{\delta}^*, 0)^T$, where $\boldsymbol{\delta}^* = (\delta_1, \ldots, \delta_K)^T$. Since the second derivative of a natural cubic spline is taken to be zero at the boundaries, $f''(a) = \delta_0 = 0$ and $f''(b) = \delta_{K+1} = 0$. As a consequence, the elements in the matrices $\boldsymbol{B}$ and $\boldsymbol{G}$ in (4.8) are defined by

$$\boldsymbol{G}_{j,j} = \frac{1}{h_j},$$
$$\boldsymbol{G}_{j,j+1} = -\frac{1}{h_{j+1}} - \frac{1}{h_j},$$
$$\boldsymbol{G}_{j,,j+2} = \frac{1}{h_{j+1}},$$

for $j = 1, \ldots, K$ and

$$\begin{aligned}
\boldsymbol{B}_{j,j} &= \frac{1}{3}(h_j + h_{j+1}), \quad j = 1, \ldots, K, \\
\boldsymbol{B}_{j,j+1} &= \frac{1}{6}h_{j+1}, \quad j = 1, \ldots, K-1, \\
\boldsymbol{B}_{j+1,j} &= \frac{1}{6}h_{j+1}, \quad j = 1, \ldots, K-1.
\end{aligned}$$

If the symmetric $K \times K$ matrix $\boldsymbol{B}$ is invertible then $\boldsymbol{\delta}^* = (\delta_1, \ldots, \delta_K)^T$ can be written in terms of $\boldsymbol{\gamma}$ as

$$\boldsymbol{\delta^*} = \boldsymbol{B}^{-1}\,\boldsymbol{G}\,\boldsymbol{\gamma}\,.$$

Therefore, $\delta_j$ and $\delta_{j+1}$ in model (4.5) can be written in terms of $\boldsymbol{\gamma}$. As a result, the only remaining parameters that need to be estimated to get a cubic spline are the $\gamma_j$'s, $j = 0, \ldots, K+1$, which can be interpreted as the function value at the knots $\kappa_j$. Therefore, the new representation of model (4.5) is

$$\begin{aligned}
f_{[j+1]}(x) \;=\; & \frac{\kappa_{j+1} - x}{h_j}\gamma_j + \frac{x - \kappa_j}{h_j}\gamma_{j+1} + \frac{1}{6}\,\boldsymbol{f}_j\,\boldsymbol{\gamma}\left(\frac{(\kappa_{j+1} - x)^3}{h_j} - h_j(\kappa_{j+1} - x)\right) \\
& + \frac{1}{6}\,\boldsymbol{f}_{j+1}\,\boldsymbol{\gamma}\left(\frac{(x - \kappa_j)^3}{h_j} - h_j(x - \kappa_j)\right),
\end{aligned} \tag{4.9}$$

where $\boldsymbol{F} = \begin{pmatrix} 0 \cdots 0 \\ \boldsymbol{B}^{-1}\boldsymbol{G} \\ 0 \cdots 0 \end{pmatrix}$ is a $(K+2) \times (K+2)$ matrix and $\boldsymbol{f}_j$ denotes the $j$-th row of $\boldsymbol{F}$.

Since in (4.9) $f_{[j+1]}(x)$ is linear in $\boldsymbol{\gamma}$, $f(x)$ is linear in $\boldsymbol{\gamma}$ and can be written as $f(x) = \sum_{k=0}^{K+1} \gamma_k b_k(x)$. In addition, the basis functions can be derived from (4.9) and can thereby be computed. In Figure 4.6 the basis functions of a cubic spline ($m = 3$) for $K = 2$ are plotted. The position of the two inner knots and the border is represented by dashed grey lines, while the basis functions are shown as continuous black lines. In the following R-code the code to generate the plots in Figure 4.6 is added.

The assumption of an additive model of the form (4.1) with normally distributed responses is equivalent to a linear regression model. Therefore, it can be easily estimated with the use of the R command `lm`. The following R-code illustrates the generation of the model matrix and the estimation of the parameters. At first the matrices $\boldsymbol{G}$, $\boldsymbol{B}$ and $\boldsymbol{F}$ are build. Thereafter, in two loops the model matrix $\boldsymbol{X}$ can be generated. While in the first loop the index `i` represents the current interval $[\kappa_{i-1}, \kappa_i]$, the index `k` in the second loop stands for the index of the parameter $\gamma_k$ corresponding to the basis function $b_k(\texttt{temp})$. In other words, the index `i` of the first loop corresponds to a set of rows of $\boldsymbol{X}$, while the index `k` denotes the current column of $\boldsymbol{X}$. Finally, the parameters can be estimated using `lm`. In the last lines of the R-code below the code to generate the plots in Figure 4.6 and Figure 4.7 is given.

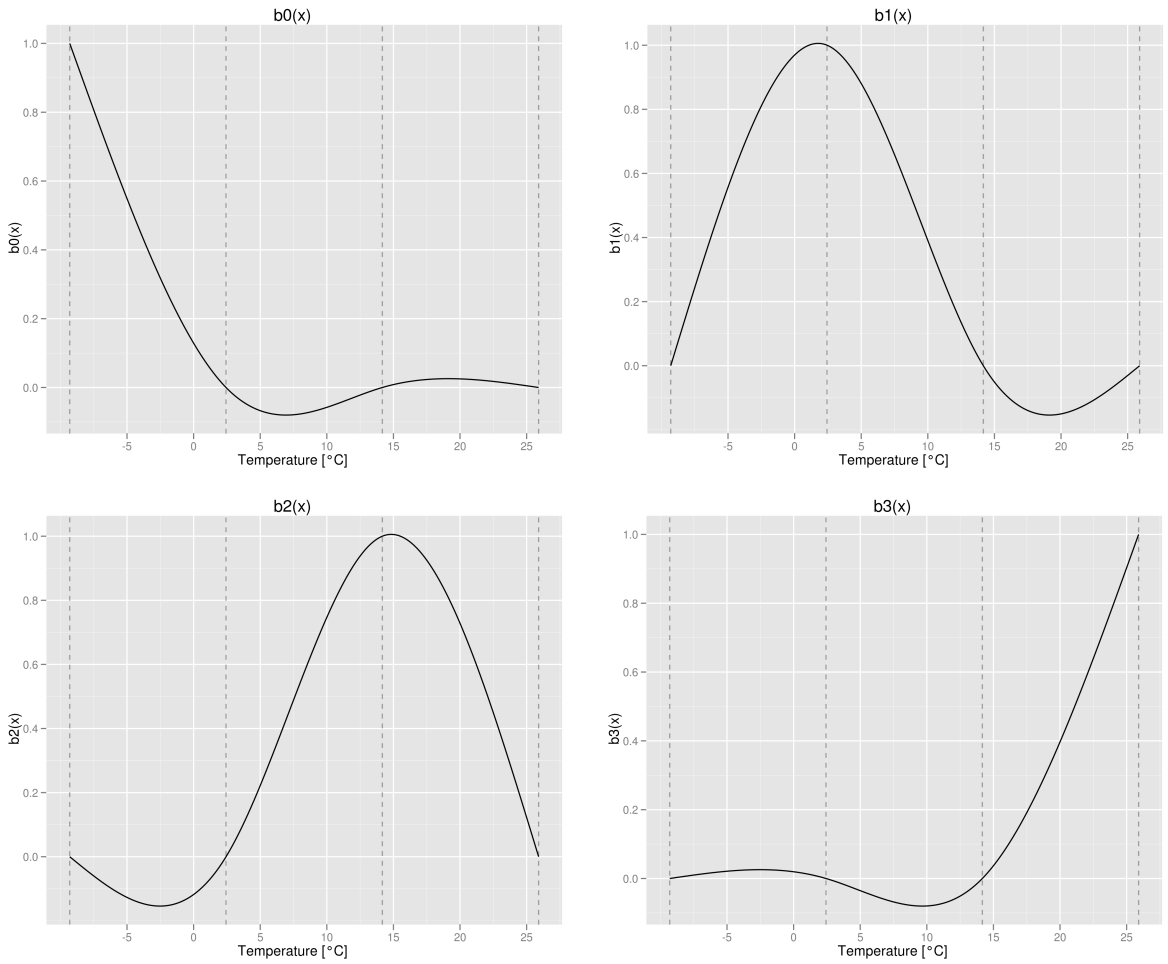


Figure 4.6.: Plot of the four basis functions of the cubic spline ($m = 3$) resulting from $K = 2$.

```
K<-5
knots<-seq(min(temp),max(temp),length.out=(K+2))
h<-knots[-1]-knots[-(K+2)] #difference between knots

G<-matrix(0,K,K+2)
B<-matrix(0,K,K)
for(i in 1:(K)) {
G[i,i]<-1/h[i]
G[i,i+1]<--1/h[i+1]-1/h[i]
G[i,i+2]<-1/h[i+1]

B[i,i]<-1/3*(h[i]+h[i+1])
if(K+2-3-i>=0) {
B[i,i+1]<-1/6*h[i+1]
```

```
B[i+1,i]<-1/6*h[i+1] }
}

F<-rbind(rep(0,K+2),solve(B)%*%G,rep(0,K+2))

X<-matrix(0,length(temp),K+2); count<-1

for(i in 1:(K+1)) { #define model matrix; i... current interval
   tempi<-temp[temp>=knots[i] & temp<knots[i+1]]
   if(i==(K+1)) tempi<-temp[temp>=knots[i] & temp<=knots[i+1]]
   Xi<-matrix(NA,length(tempi),K+2)

   for(k in 1:(K+2)) {               # k... current column of X
       Xi[,k]<-1/6*F[i,k]*((knots[i+1]-tempi)^3/h[i] -
                           h[i]*(knots[i+1]-tempi))+
               1/6*F[i+1,k]*((tempi-knots[i])^3/h[i] -
                            h[i]*(tempi-knots[i]))
       if(i==k) Xi[,k]<-Xi[,k]+(knots[i+1]-tempi)/h[i]
       if((i+1)==k) Xi[,k]<-Xi[,k]+(tempi-knots[i])/h[i]
   }
   X[count:(count+length(tempi)-1),]<-Xi
   count<-count+length(tempi)
}

mod.cubic<-lm(max.flow~X-1)

#plot basis functions
data_cubic<-data.frame(temp, X)
colnames(data_cubic)<-c("temp",paste("b",1:(K+2),sep=""))
knots<-seq(min(temp),max(temp),length.out=(K+2))

line_knots<-geom_vline(xintercept=knots,linetype=2,alpha=1/3)
qplot(temp,b1,data=data_cubic,geom="line") +line_knots


#plot fit
p +geom_line(aes(x=temp,y=X%*%coef(mod.cubic)),colour="red")
  +line_knots
```

The fits resulting from this model for $K = 1, 5, 10, 15$ are shown in Figure 4.7. Again the number of knots $K$ strongly influences the fit. One can observe that a large number of knots leads to a wiggly fit (lower right plot) while a too small number results in an unsatisfying fit (upper left plot).

The fact that we have to choose again the number of knots and thereby the degree

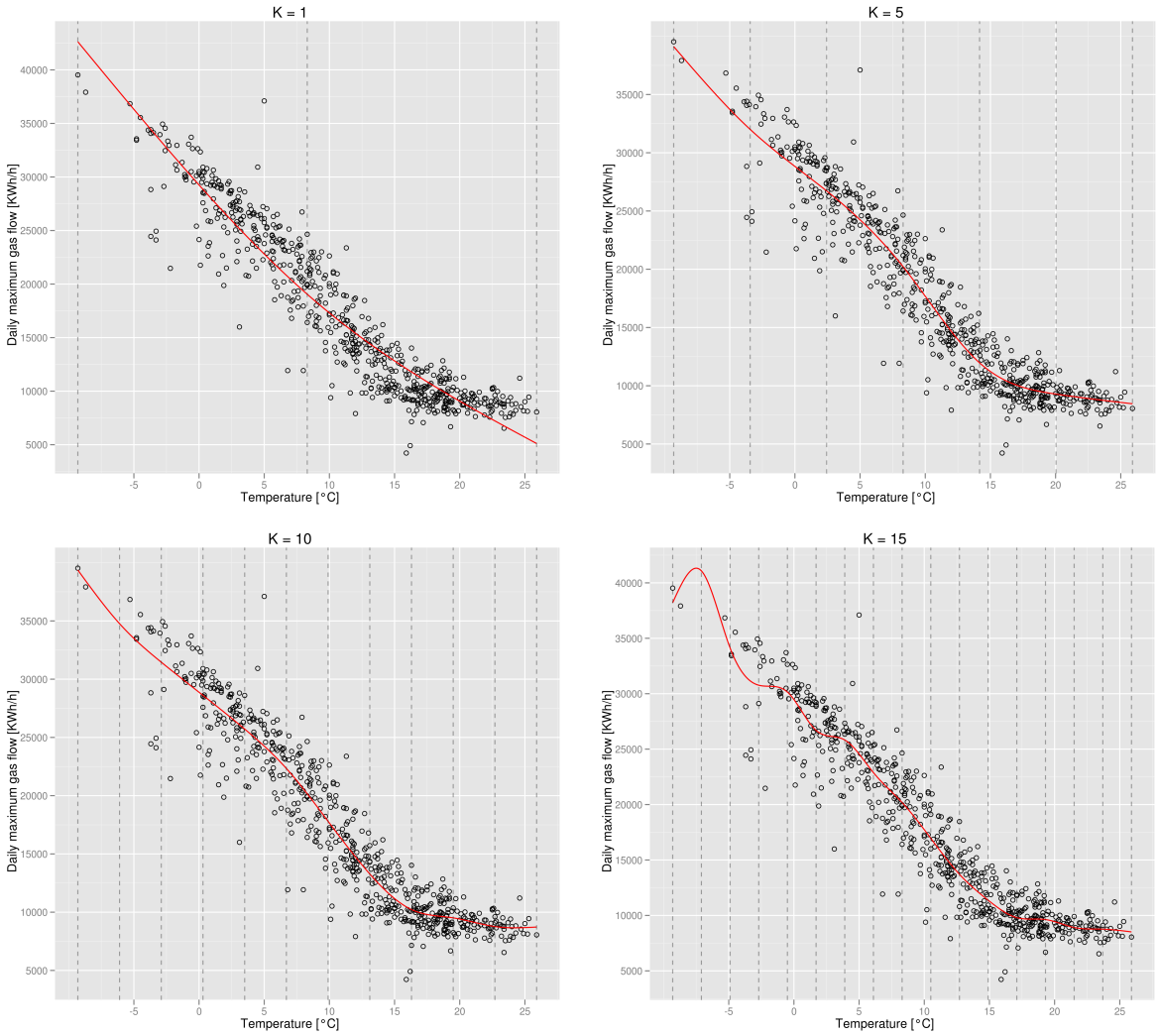


Figure 4.7.: Smooth fit resulting from cubic spline basis functions for $K = 1, 5, 10, 15$ $(m = 3)$.

of smoothness turns out to be a disadvantage of this method. On the other hand, the choice of the parameters as $\gamma_j = f(\kappa_j)$ is useful because it allows for an interpretation of them.

### B-Splines

Another possible choice of basis functions are B-splines. They will turn out to be very useful in the context of penalized splines, which we will discuss later in Section 4.2. But first of all we need to define them, which is easily done in a recursive manner. Starting

with a B-spline of order $m = 0$

$$B_j^0(x) = \begin{cases} 1 & \kappa_j \leq x \leq \kappa_{j+1} \\ 0 & \text{otherwise} \end{cases} \qquad j = 0, \ldots, K,$$

B-splines of higher order are defined by

$$B_j^m(x) = \frac{x - \kappa_j}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) + \frac{\kappa_{j+m+1} - x}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x),$$

where $j = -m, \ldots, K$. This index is a little bit peculiar because we also need B-splines which start outside of the $x$ range but are partially in the observed interval.

Similar as before the smooth function $f(x)$ with basis functions $B_j^m(x)$ is given by

$$f(x) = \sum_{j=-m}^{K} \gamma_j B_j^m(x).$$

For a better understanding B-splines of order $m = 0, 1, 2, 3$ are shown in Figure 4.8. While the basis functions with $m = 0$ are step functions on the intervals $[\kappa_j, \kappa_{j+1}]$, higher order B-splines are linear, quadratic or cubic functions for example. The knots $\kappa_j$, $j = 0, \ldots, K+1$, in Figure 4.8 are again represented by dashed grey lines. We observe that a B-spline of order $m$ is nonzero over an interval of $m + 2$ knots. In addition, the number of B-splines on the data range is given by $q = K + m + 1$, while the number of needed knots is $K + 2m + 2$. This last number is larger than $K$ because knots outside the data range are needed to construct the B-splines which are only partially inside the observed $x$-range and partially beyond. Therefore, to construct B-splines $B_j^m(x)$, $j = -m, \ldots, K$, of degree $m$ on $[a, b]$ a vector of knots $\boldsymbol{\kappa} = (\kappa_{-m}, \ldots, \kappa_{K+m+1})$ is used.

Another useful attribute of B-splines is that their derivatives can be written in form of differences. We can show that the first derivative of a B-spline of order $m$ is

$$\frac{\partial B_j^m(x)}{\partial x} = m \left( \frac{1}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) - \frac{1}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x) \right).$$

The proof is given in Appendix A. As a result,

$$\frac{\partial}{\partial x} \sum_{j=-m}^{K} \gamma_j B_j^m(x) = m \sum_{j=-m}^{K} \frac{\gamma_j - \gamma_{j-1}}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x). \tag{4.10}$$

In the command `splineDesign` of the package `splines` in R the number of coefficients in each piecewise polynomial segment defines the degree of a spline. This definition of the degree of a spline differs from our definition (degree of the highest polynomial) by one. For example, in our definition a cubic spline has degree three, while in `splineDesign` it has degree four. As a consequence in the following R-code `ord`$= m + 1 = 4$ to get a cubic spline.

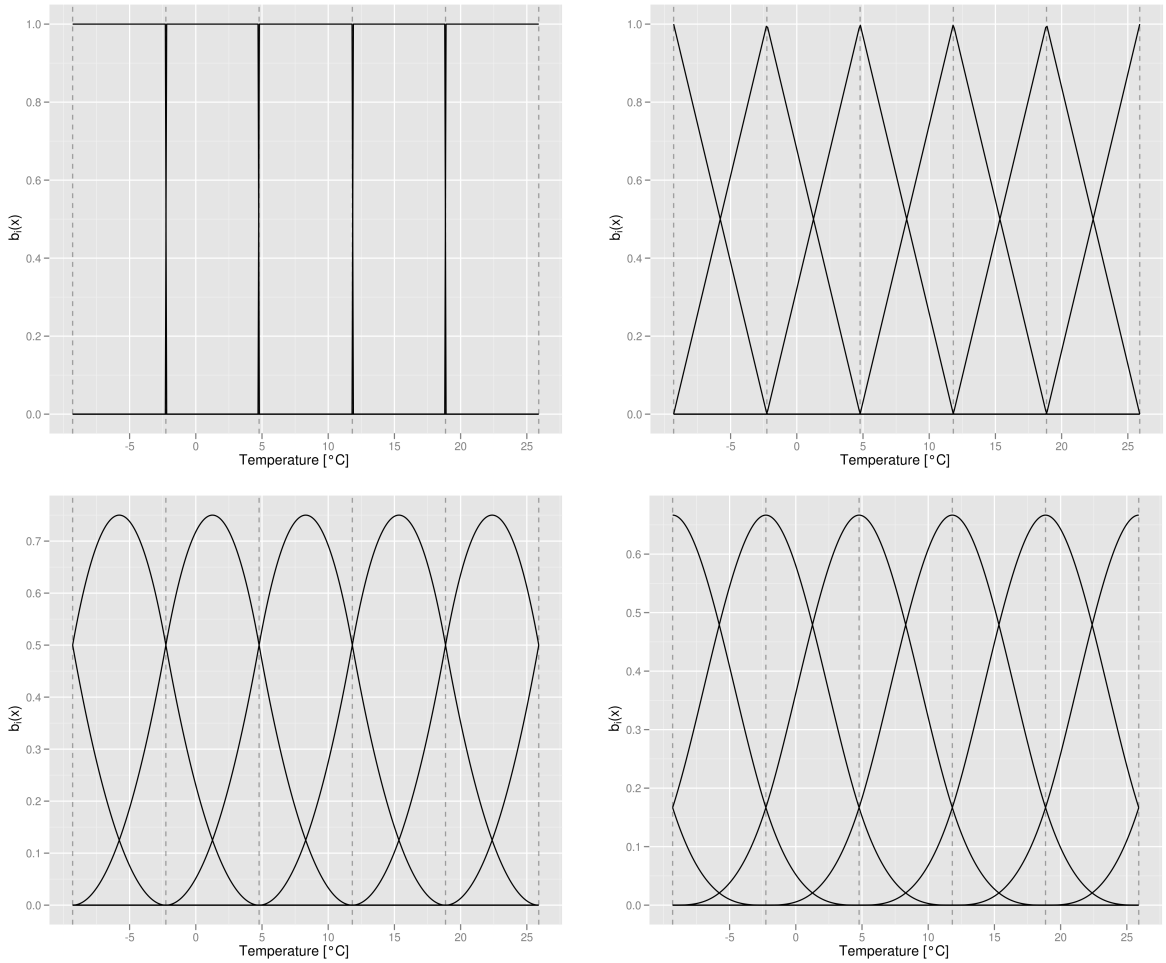


Figure 4.8.: Plot of the B-spline functions resulting from $K = 4$ and $m = 0, 1, 2, 3$.

In the R-code below at first the order of the B-splines and the number of inner knots are defined. Thereafter the length of the distance between the equidistant knots is calculated and later used by the function `knots` to calculate them. A definition of this function is given at the end of the R-code. The function `splineDesign` of the package `splines` constructs the model matrix consisting of B-spline basis functions. As arguments of this function we provide the knots we want to use, the values $x$ at which we want to evaluate the B-splines and the order of the B-spline functions, which differs from $m$ by one.

Assuming an additive model of the form (4.1) with normally distributed responses, we can again reduce our model to a linear regression model. Therefore, we are able to estimate it with the R-function `lm`. Thereafter, the code for the plots of the fitted values is added and its result shown in Figure 4.9.

```
m<-3; K<-4

Bs<-splineDesign(knots((-m):(K+m+1),temp,K),temp,ord=m+1)
```

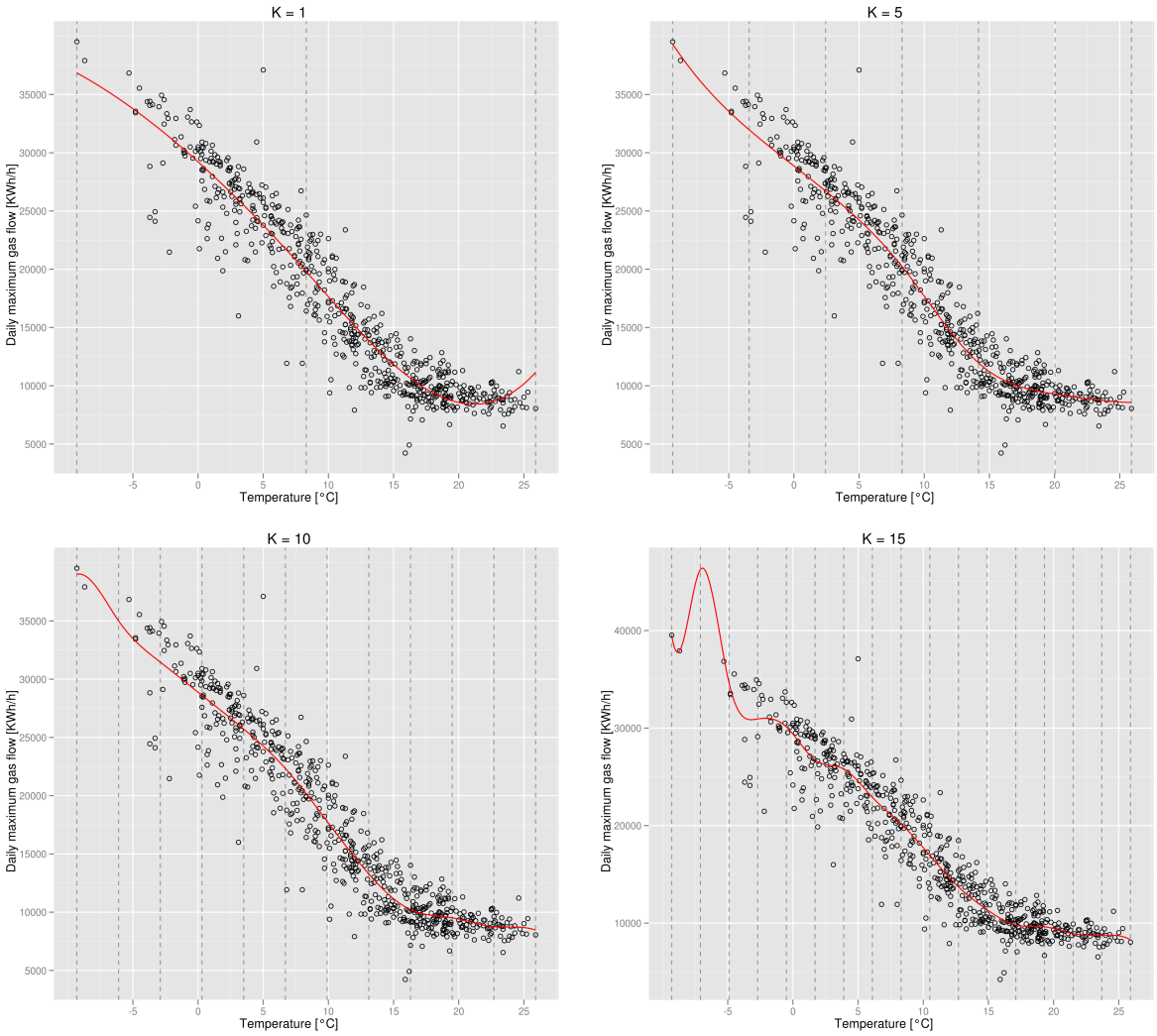


Figure 4.9.: Smooth fit resulting from cubic B-splines for $K = 1, 5, 10, 15$.

```
mod.bspline<-lm(max.flow~Bs-1)

line_knots<-geom_vline(xintercept=knots(0:(K+1),temp,K),
   linetype=2,alpha=1/3)
p +geom_line(aes(x=temp,y=Bs%*%coef(mod.bspline)),colour="red")
  +line_knots                              #plot of data, knots, fit

#function knots
knots<-function(i,x,K) {
  s<-(max(x)-min(x))/(K+1)
  k<-min(x)+i*s
  k
}
```

The fits in Figure 4.9 show a similar behaviour as the fits resulting from the basis functions which we discussed previously. For $K = 1$ the fit lacks flexibility and is not able to describe the data properly. While the fit for $K = 5$ seems fairly okay, the fits for $K = 10$ and $K = 15$ are too flexible and do not represent the overall trend, especially at intervals with less observations.

The disadvantage of B-splines turns out to be the same as we had before. The influence of the choice of knots on the fit proofs to be a problem. Therefore, in the next subsection a solution of this problem is discussed.

The greatest advantage of B-splines is their use for P-splines, which will be introduced in the next subsection. In addition, they are easy to understand, flexible and the parameters can be interpreted as the function values at the apex of a basis function. More information on B-splines can be found in Eilers and Marx (1996).

## 4.2. Penalized Splines

The issue how to choose the number of knots and thereby the degree of smoothness addressed in the last subsection is picked up now and a possible solution is discussed. The idea is to allow a high number of knots and thereby a high degree of flexibility but to control the degree of smoothness with a penalty term.

If we use B-splines as basis functions and a penalized approach as will be discussed in this subsection, then one talks of penalized splines, or P-splines in short. As a result, P-splines are basically B-splines but their parameters $\gamma_j$, $j = 0, \ldots, q-1$, are estimated using a penalized least squares approach. While Wood (2006a) uses the integral of the squared second derivative of $f(x)$ as a penalty, Eilers and Marx (1996) suggest to use parameter differences as penalty. The differences between these two penalties will be addressed and a comparison of the resulting fits based on these penalties is provided.

### 4.2.1. Model Estimation

In the following, a penalized least squares approach is introduced and discussed. During this subsection we consider an additive model as in (4.1) with normally distributed responses. But contrary to the methods in Chapter 2, we now estimate the model using the penalized least squares approach

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 - \lambda \int_a^b f''(x)^2 \mathrm{d}x, \tag{4.11}$$

where $\lambda \geq 0$ controls the degree of smoothness.

While the first part is a usual sum of squares, the second part constitutes a penalty to assure a certain degree of smoothness. All in all, in a penalized least squares approach we want to find $f$ that minimizes (4.11) for a given value of $\lambda$.

If we assume that $f$ can be written as proposed in model (4.2), then the second

derivative of $f$ has the form

$$f''(x) = \sum_{j=0}^{q-1} \gamma_j b_j''(x) = \boldsymbol{\gamma}^T \mathbf{a}(x),$$

where $a_j(x) = b_j''(x)$, and we get

$$\int_a^b f''(x)^2 \mathrm{d}x = \int_a^b \boldsymbol{\gamma}^T \mathbf{a}(x)\mathbf{a}^T(x)\,\boldsymbol{\gamma}\,\mathrm{d}x = \boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma}, \tag{4.12}$$

where $\boldsymbol{S} = \int_a^b \mathbf{a}(x)\mathbf{a}^T(x)\mathrm{d}x$.

Therefore, (4.11) can be written as

$$\mathcal{S}_q = \| \boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\gamma} \|^2 + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma},$$

where

$$\boldsymbol{x}_i^T\,\boldsymbol{\gamma} = \sum_{j=0}^{q-1} \gamma_j b_j(x_i) = f(x_i) \quad \text{for } i = 1, \ldots, n,$$

with $\boldsymbol{x}_i = (b_0(x_i), \ldots, b_{q-1}(x_i))^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$.

To minimize $\mathcal{S}_q$ with respect to $\boldsymbol{\gamma}$, we calculate the first derivative

$$\frac{\partial \mathcal{S}_q}{\partial\,\boldsymbol{\gamma}} = -2\,\boldsymbol{X}^T\,(\boldsymbol{y} - \boldsymbol{X}\,\boldsymbol{\gamma}) + 2\lambda\,\boldsymbol{S}\,\boldsymbol{\gamma}\,.$$

Setting it equal to zero leads to

$$-\,\boldsymbol{X}^T\,\boldsymbol{y} + \boldsymbol{X}^T\,\boldsymbol{X}\,\hat{\boldsymbol{\gamma}} + \lambda\,\boldsymbol{S}\,\hat{\boldsymbol{\gamma}} = \boldsymbol{0}\,.$$

As a consequence, the least squares estimator of $\boldsymbol{\gamma}$ is explicitly given by

$$\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{X}^T\,\boldsymbol{X} + \lambda\,\boldsymbol{S}\right)^{-1}\,\boldsymbol{X}^T\,\boldsymbol{y}\,. \tag{4.13}$$

Note that the estimator in (4.13) is similar to the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in a linear regression model, see (2.2). The only difference is the extra term $\lambda\,\boldsymbol{S}$, which is due to the penalization.

As a result, an estimator of the mean $\boldsymbol{\mu}$ is given by

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\,\hat{\boldsymbol{\gamma}} = \boldsymbol{X}\left(\boldsymbol{X}^T\,\boldsymbol{X} + \lambda\,\boldsymbol{S}\right)^{-1}\,\boldsymbol{X}^T\,\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{y},$$

where the so called influence matrix $\boldsymbol{A}$ is defined as

$$\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}^T\,\boldsymbol{X} + \lambda\,\boldsymbol{S}\right)^{-1}\,\boldsymbol{X}^T\,. \tag{4.14}$$

This influence matrix is similar to the hat matrix $\boldsymbol{H}$ in Chapter 2. The only difference lies again in the term $\lambda\,\boldsymbol{S}$. Thus, the influence matrix is usually not idempotent, in contrast to the hat matrix for linear models.

The integral of the squared second derivative of $f(\cdot)$ is a common penalty, see Hastie and Tibshirani (1990) or Wood (2006a). One advantage of this choice of penalty is that the function $f(x)$ which minimizes (4.11) is known, see Theorem 1.

**Theorem 1.** *For all continuous functions with continuous first and integrable second derivative on the x range $[a, b]$ the cubic spline $g(x)$ is the function minimizing*

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_a^b g''(x)^2 \mathrm{d}x, \tag{4.15}$$

*where $\lambda \geq 0$ is fixed.*

*Proof.* This proof is taken from Wood (2006a). If any other continuous function $h(x)$ with continuous first and integrable second derivative minimizes equation (4.15), it is possible to choose a cubic spline $g(x)$ which interpolates $h(x)$ in the points $(x_i, h(x_i))$. Therefore, the first part of (4.15) is the same for $h(x)$ and $g(x)$, i.e.

$$\sum_{i=1}^{n} (y_i - h(x_i))^2 = \sum_{i=1}^{n} (y_i - g(x_i))^2 .$$

In regard of the second part, a new function $d(x) = h(x) - g(x)$ is considered

$$\begin{aligned} \int_a^b h''(x)^2 \mathrm{d}x &= \int_a^b \left(d''(x) + g''(x)\right)^2 \mathrm{d}x \\ &= \int_a^b d''(x)^2 \mathrm{d}x + 2\int_a^b d''(x) g''(x) \mathrm{d}x + \int_a^b g''(x)^2 \mathrm{d}x. \end{aligned}$$

For the second term of the last result partial integration and the assumption $g''(a) = g''(b) = 0$ ensure that

$$\begin{aligned} \int_a^b d''(x) g''(x) \mathrm{d}x &= \underbrace{g''(x) d'(x)|_a^b}_{=0} - \int_a^b g'''(x) d'(x) \mathrm{d}x \\ &= -\int_a^b g'''(x) d'(x) \mathrm{d}x. \end{aligned}$$

Since the third derivative of the cubic spline $g(x)$ is a step function, it follows that

$$\begin{aligned} -\int_a^b g'''(x) d'(x) \mathrm{d}x &= -\sum_{i=1}^{n-1} g'''(x_i^*) \int_{x_i}^{x_{i+1}} d'(x) \mathrm{d}x \\ &= -\sum_{i=1}^{n-1} g'''(x_i^*) \left(d(x_{i+1}) - d(x_i)\right) = 0, \end{aligned}$$

where $x_i^*$ is an element in the interval $(x_i, x_{i+1})$, and the last result is due to the fact that for each $x_i$ the functions $h(x)$ and $g(x)$ are the same, since $g(x)$ interpolates $h(x)$ in $(a, \ldots, b)$, and therefore $d(x_i)$ is zero for $i = 1, \ldots, n$.

In total, for the second part of (4.15) it follows that

$$\begin{aligned}\int_a^b h''(x)^2 \mathrm{d}x &= \int_a^b d''(x)^2 \mathrm{d}x + \int_a^b g''(x)^2 \mathrm{d}x \\ &\geq \int_a^b g''(x)^2 \mathrm{d}x.\end{aligned}$$

Summarizing, the first term in (4.15) is equal for $h$ and $g$, while the second term, namely $\lambda \int_a^b (g''(x))^2 \mathrm{d}x$, is smaller or equal for $g$ compared to $h$. Therefore, if $h$ minimizes (4.15), $g$ minimizes it too. □

### 4.2.2. Effective Degrees of Freedom and Residual Variance

Now the question how many degrees of freedom an additive model has is addressed. The effective degrees of freedom are a measure for the flexibility of a model or the wiggliness of the fit. For example, if $\lambda = 0$, then the penalized least squares criterion (4.11) is minimal if $f(x_i) = y_i$ and therefore the fit interpolates the data points. This is only the case if $q$ is sufficiently large as it is assumed in case of P-splines. On the other hand, if $\lambda$ tends towards infinity, then the second part of (4.11) gets large unless $f''(x) = 0$, resulting in a linear fit. As a result, the maximal effective degrees of freedom corresponding to the maximal degree of flexibility are equal to the number of parameters. While the effective degrees of freedom corresponding to a linear fit and thereby little flexibility constitute the minimum.

Taking the linear model case as an example, the degrees of freedom can be defined as the trace of the hat matrix $\boldsymbol{H}$ or in this case the influence matrix $\boldsymbol{A}$

$$p = \mathrm{tr}(\boldsymbol{A}),$$

while the resiudal degrees of freedom are $n - p = n - \mathrm{tr}(\boldsymbol{A})$.

If we are interested in the degrees of freedom of the parameter $\boldsymbol{\gamma}$, then we define

$$\boldsymbol{P} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X}^T,$$

where $\hat{\boldsymbol{\gamma}} = \boldsymbol{P}\,\boldsymbol{y}$ and $\boldsymbol{A} = \boldsymbol{X}\,\boldsymbol{P}$. As a result, the trace of $\boldsymbol{A}$ can be written as

$$\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{X}\,\boldsymbol{P}) = \sum_{i=1}^{n} (\boldsymbol{P}\,\boldsymbol{X})_{i,i}.$$

As a consequence, the diagonal element $(\boldsymbol{P}\,\boldsymbol{X})_{i,i}$ describes the effective degrees of freedom of the $i$-th parameter. In other words, the diagonal elements of the matrix

$$\boldsymbol{P}\,\boldsymbol{X} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X}^T \boldsymbol{X}$$

are the effective degrees of freedom of the respective parameters in the additive model case.

Next, we want to introduce an estimator for the response variance $\sigma^2$. As in case of a linear model an estimator for $\sigma^2$ is given by the residual sum of squares divided by the residual degrees of freedom. Adapting this approach to an additive model leads to

$$\hat{\sigma}^2 = \frac{\| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2}{n - \mathrm{tr}(\boldsymbol{A})}. \tag{4.16}$$

Unfortunately, the estimator (4.16) is not unbiased. This can be easily shown by looking at

$$\begin{aligned}
\mathbb{E}\left[\| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2\right] &= \mathbb{E}\left[\|(\boldsymbol{\mu} + \boldsymbol{\varepsilon}) - \boldsymbol{A}(\boldsymbol{\mu} + \boldsymbol{\varepsilon})\|^2\right] \\
&= \mathbb{E}\left[\|(\boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{\mu}) + (\boldsymbol{\varepsilon} - \boldsymbol{A}\,\boldsymbol{\varepsilon})\|^2\right] \\
&= \| \boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{\mu} \|^2 + \mathbb{E}\left[(\boldsymbol{\varepsilon} - \boldsymbol{A}\,\boldsymbol{\varepsilon})^T(\boldsymbol{\varepsilon} - \boldsymbol{A}\,\boldsymbol{\varepsilon}) + 2(\boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{\mu})^T(\boldsymbol{\varepsilon} - \boldsymbol{A}\,\boldsymbol{\varepsilon})\right] \\
&= \| \boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{\mu} \|^2 + \mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{\varepsilon} - 2\,\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\,\boldsymbol{A}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon}\right] \\
&= \| \boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{\mu} \|^2 + n\sigma^2 - 2\,\mathbb{E}\left[\mathrm{tr}(\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon})\right] + \mathbb{E}\left[\mathrm{tr}(\boldsymbol{\varepsilon}^T\,\boldsymbol{A}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon})\right] \\
&= \| \boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{\mu} \|^2 + \sigma^2\left(n - 2\mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{A}^T\,\boldsymbol{A})\right),
\end{aligned}$$

where we assume that $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ with $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\,\boldsymbol{I}$ and where we use

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{\mu}\right] &= \mathbb{E}\left[\boldsymbol{\varepsilon}^T\right]\boldsymbol{\mu} = 0, \\
\mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\mu}\right] &= \mathbb{E}\left[\boldsymbol{\varepsilon}^T\right]\boldsymbol{A}\,\boldsymbol{\mu} = 0, \\
\mathbb{E}\left[\mathrm{tr}(\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon})\right] &= \mathbb{E}\left[\mathrm{tr}(\boldsymbol{A}\,\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^T)\right] = \mathrm{tr}\left(\boldsymbol{A}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^T\right]\right) \\
&= \mathrm{tr}(\boldsymbol{A}\,\boldsymbol{I})\,\sigma^2 = \mathrm{tr}(\boldsymbol{A})\,\sigma^2.
\end{aligned}$$

### 4.2.3. Difference Penalty

The aim of this part is to introduce a different penalty term than the integral of the squared second derivative. The difference penalty is proposed in Eilers and Marx (1996). In the following we will show that this penalty is similar to the penalty we have used in (4.11), but is computationally more efficient and simpler.

While until now the penalty was given by

$$P_1 = \lambda \int_a^b f''(x)^2 \mathrm{d}x, \tag{4.17}$$

where $\lambda$ is the respective smoothing parameter, Eilers and Marx (1996) propose a difference penalty of the form

$$P_2 = \lambda \sum_{j=k+1}^{q-1} \left(\Delta^k \gamma_j\right)^2, \tag{4.18}$$

where $\Delta^k$ describes the $k$-th difference of the parameter vector $\boldsymbol{\gamma}$. For example the first and second order differences of $\gamma_j$ are

$$\begin{aligned}
\Delta\gamma_j &= \gamma_j - \gamma_{j-1}, \\
\Delta^2\gamma_j &= \Delta(\Delta\gamma_j) = \Delta\gamma_j - \Delta\gamma_{j-1} \\
&= \gamma_j - \gamma_{j-1} - (\gamma_{j-1} - \gamma_{j-2}) = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}.
\end{aligned}$$

In addition to the smoothing parameter $\lambda$ now the order of the difference $k$ needs to be chosen too.

As Eilers and Marx (1996) show, if we consider B-splines of order three as our basis functions, then there is not a big difference between $P_1$ and $P_2$ if in the second penalty differences of order $k = 2$ are considered. If $f(x) = \sum_{j=0}^{q-1} \gamma_j B_j^3(x)$, then $P_1$ can be written as

$$P_1 = \lambda \int_a^b \left( \frac{\partial^2}{\partial x^2} \sum_{j=0}^{q-1} \gamma_j B_j^3(x) \right)^2 \mathrm{d}x.$$

Since we did already mention in (4.10) that the sum of derivatives of B-splines can be written as a sum of differences of the B-spline parameters, it follows that

$$\begin{aligned} P_1 &= \lambda \int_a^b \left( \sum_{j=0}^{q-1} \Delta^2 \gamma_j B_j^1(x) \right)^2 \mathrm{d}x, \\ &= \lambda \int_a^b \sum_{j=0}^{q-1} \sum_{i=0}^{q-1} \Delta^2 \gamma_j \Delta^2 \gamma_i B_j^1(x) B_i^1(x) \mathrm{d}x. \end{aligned}$$

Another useful property of B-splines of order $m = 1$ is that they only overlap with their direct neighbours. As a result, all cross products in the equation above are zero except for $j = i - 1$ and $j = i + 1$. Therefore,

$$\begin{aligned} P_1 &= \lambda \int_a^b \left( \sum_{j=0}^{q-1} \left( \Delta^2 \gamma_j B_j^1(x) \right)^2 + 2 \sum_{j=0}^{q-1} \Delta^2 \gamma_j \Delta^2 \gamma_{j-1} B_j^1(x) B_{j-1}^1(x) \right) \mathrm{d}x \\ &= \lambda \sum_{j=0}^{q-1} \left( \Delta^2 \gamma_j \right)^2 \int_a^b \left( B_j^1(x) \right)^2 \mathrm{d}x + 2\lambda \sum_{j=0}^{q-1} \Delta^2 \gamma_j \Delta^2 \gamma_{j-1} \int_a^b B_j^1(x) B_{j-1}^1(x) \mathrm{d}x. \end{aligned}$$

While the first term of $P_1$ is equivalent to $P_2$, the second term is the product of second order differences of neighbouring B-splines. Therefore, $P_1$ would yield to a more complex penalty than $P_2$, which can get problematic if higher orders of B-splines are considered. As a consequence, a difference penalty could yield to similar results as $P_1$ while it is computationally less expensive. An application of this theory on the gas flow data can be observed later in Chapter 5, where the two fits which result from these two penalties are compared.

Another way to write the penalty $P_2$ in (4.18) is

$$P_2 = \lambda \sum_{j=k+1}^{q-1} \left( \Delta^k \gamma_j \right)^2 = \lambda \boldsymbol{\gamma}^T \boldsymbol{D}_k^T \boldsymbol{D}_k \boldsymbol{\gamma},$$

where $\boldsymbol{D}_k$ is a $(q - k) \times q$ difference matrix. For example, a difference matrix of order

$k = 2$ has the form

$$\boldsymbol{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

As a result, we are able to write $P_2$ as

$$\begin{aligned} P_2 &= \lambda \boldsymbol{\gamma}^T \boldsymbol{D}_k^T \boldsymbol{D}_k \boldsymbol{\gamma} \\ &= \lambda \boldsymbol{\gamma}^T \boldsymbol{S} \boldsymbol{\gamma}, \end{aligned} \tag{4.19}$$

where $\boldsymbol{S} = \boldsymbol{D}_k^T \boldsymbol{D}_k$. Since this representation of $P_2$ is similar to the penalty in Subsection 4.2, the theory introduced there applies in case of the difference penalty too.

## 4.3. Smoothing Parameter $\lambda$

In Section 4.2 we could transform the problem of the choice of the number of knots to the problem of choosing the smoothing parameter $\lambda$. This section introduces criteria as to how $\lambda$ could be chosen.

In Figure 4.10 four different P-splines with four different $\lambda$ values are visualised. While the fit in the first plot is the most wiggly one, the following fits are a lot smoother. In addition, one can observe how with an increasing penalty the fit converges to a linear function.

In case of a large penalty value, a P-spline is dominated by the penalty term and therefore converges to a polynomial function of degree $k - 1$, where $k$ describes the difference order. This is similar to the case of a derivative penalty, where a large value of $\lambda$ forces $f''(x) = 0$, and therefore $f(x)$ results in a linear function. In that sense, a large value of $\lambda$ ensures that $\Delta^k \gamma_j$ tends towards zero for all $j$ and the result is a polynomial of degree $k - 1$. In Figure 4.11 one can observe the same plots if the order of the differences is $k = 3$.

Figure 4.11 shows the same behaviour as Figure 4.10. But with increasing $\lambda$ the fit now converges to a quadratic polynomial function, because $k - 1 = 2$.

### 4.3.1. Unbiased Risk Estimator

The first criterion for the choice of $\lambda$ which we discuss is the unbiased risk estimator or UBRE. The idea is to choose $\lambda$ so that $\hat{\boldsymbol{\mu}}$ is close to the true parameter $\boldsymbol{\mu}$. Therefore, we take a look at the mean squared error $MSE$ to derive the UBRE score.

In case of an additive model the mean squared error takes the form

$$MSE = \mathbb{E}\left[\frac{1}{n} \| \boldsymbol{\mu} - \boldsymbol{X}\, \hat{\boldsymbol{\gamma}} \|^2\right] = \frac{1}{n}\, \mathbb{E}\left[\| \boldsymbol{y} - \boldsymbol{A}\, \boldsymbol{y} \|^2\right] - \sigma^2 + \frac{2\sigma^2}{n} \mathrm{tr}(\boldsymbol{A}), \tag{4.20}$$

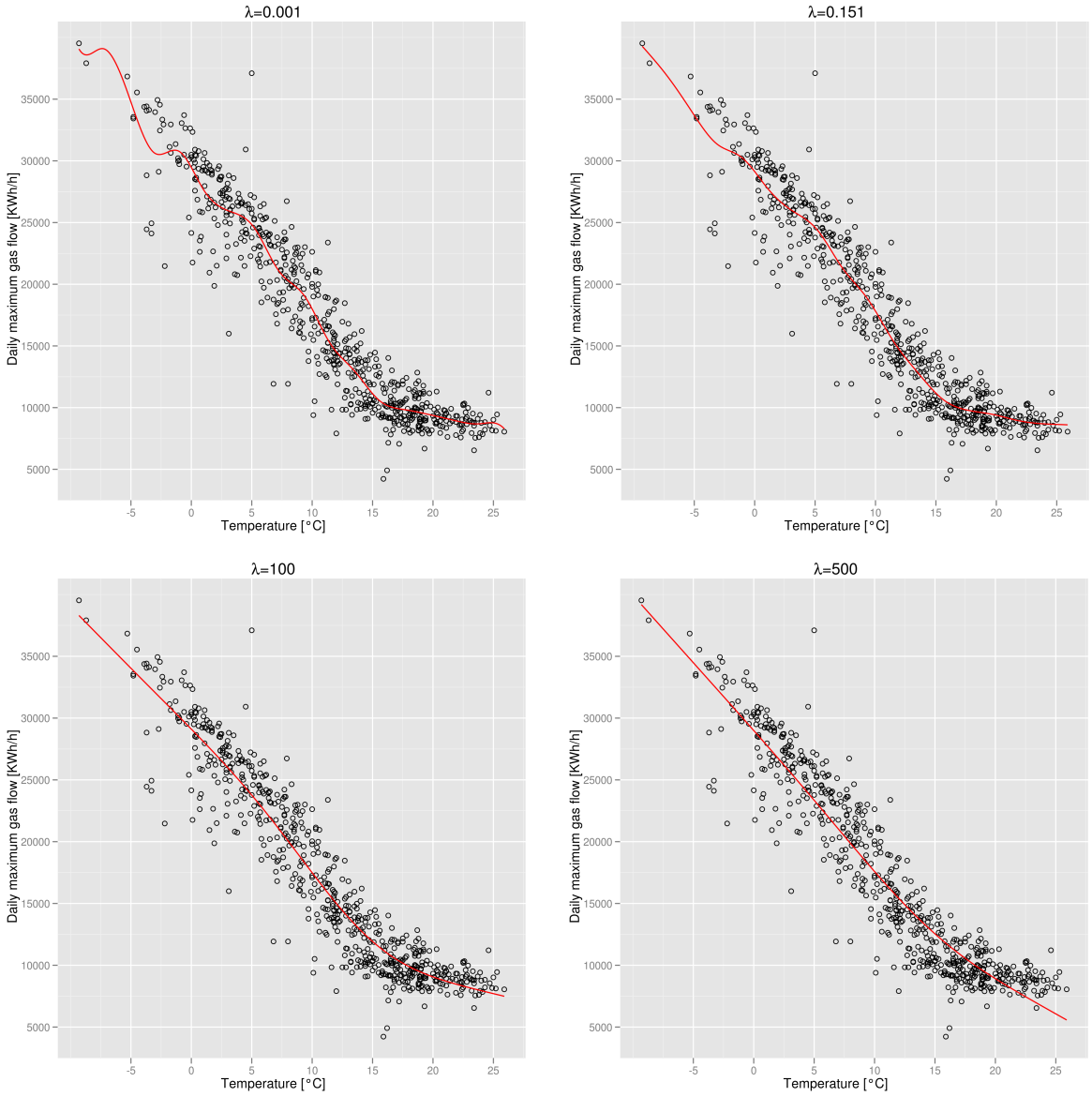


Figure 4.10.: P-splines with difference order $k = 2$, $q = 20$ and smoothing parameter $\lambda = 0.001, 0.1515, 100, 500$.

where the last result follows from

$$\begin{aligned}
\| \boldsymbol{\mu} - \boldsymbol{X}\,\hat{\boldsymbol{\gamma}} \|^2 &= \| \boldsymbol{\mu} - \boldsymbol{A}\,\boldsymbol{y} \|^2 \\
&= \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} - \boldsymbol{\varepsilon} \|^2 \\
&= \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2 + \boldsymbol{\varepsilon}^T\,\boldsymbol{\varepsilon} - 2\,\boldsymbol{\varepsilon}^T\,(\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y}) \\
&= \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2 + \boldsymbol{\varepsilon}^T\,\boldsymbol{\varepsilon} - 2\,\boldsymbol{\varepsilon}^T\,((\boldsymbol{\mu} + \boldsymbol{\varepsilon}) - \boldsymbol{A}(\boldsymbol{\mu} + \boldsymbol{\varepsilon})) \\
&= \| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2 - \boldsymbol{\varepsilon}^T\,\boldsymbol{\varepsilon} - 2\,\boldsymbol{\varepsilon}^T\,\boldsymbol{\mu} + 2\,\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\mu} + 2\,\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon},
\end{aligned}$$

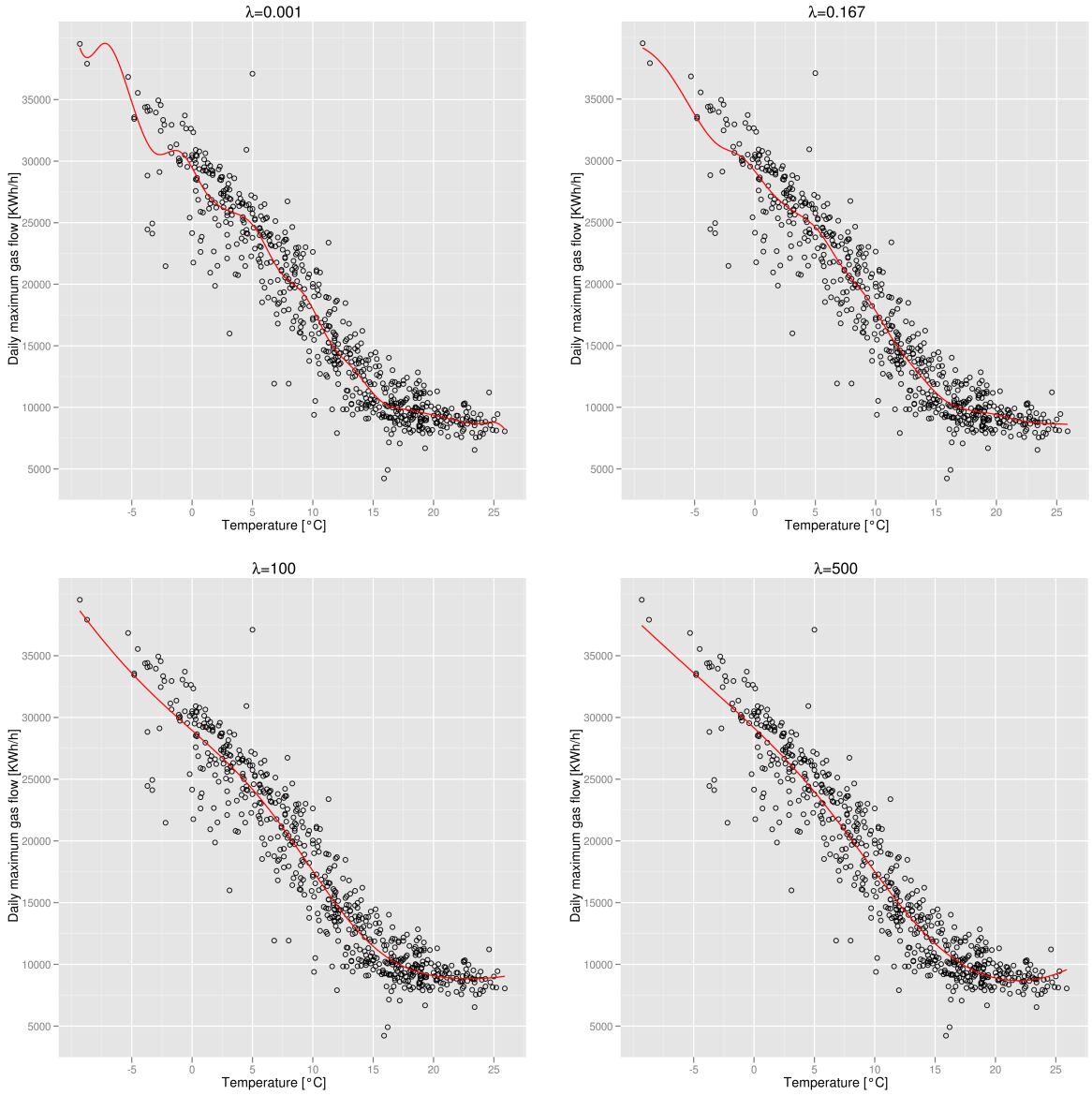


Figure 4.11.: P-splines with difference order $k = 3$, $q = 20$ and smoothing parameter $\lambda = 0.001, 0.1515, 100, 500$.

where $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ with $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\text{var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$. Taking the mean leads to

$$\begin{aligned}
\mathbb{E}\left[\| \boldsymbol{\mu} - \boldsymbol{X}\,\hat{\gamma} \|^2\right] &= \mathbb{E}\left[\| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2\right] - \mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{\varepsilon}\right] - 2\,\mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{\mu}\right] \\
&\quad + 2\,\mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\mu}\right] + 2\,\mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon}\right] \\
&= \mathbb{E}\left[\| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2\right] - n\sigma^2 + 2\,\mathbb{E}\left[\text{tr}(\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon})\right] \\
&= \mathbb{E}\left[\| \boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y} \|^2\right] - n\sigma^2 + 2\text{tr}(\boldsymbol{A})\sigma^2,
\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{\mu}\right] &= \mathbb{E}\left[\boldsymbol{\varepsilon}^T\right]\boldsymbol{\mu} = 0,\\ \mathbb{E}\left[\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\mu}\right] &= \mathbb{E}\left[\boldsymbol{\varepsilon}^T\right]\boldsymbol{A}\,\boldsymbol{\mu} = 0,\\ \mathbb{E}\left[\mathrm{tr}(\boldsymbol{\varepsilon}^T\,\boldsymbol{A}\,\boldsymbol{\varepsilon})\right] &= \mathbb{E}\left[\mathrm{tr}(\boldsymbol{A}\,\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^T)\right] = \mathrm{tr}\left(\boldsymbol{A}\,\mathbb{E}\left[\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^T\right]\right)\\ &= \mathrm{tr}(\boldsymbol{A}\,\boldsymbol{I})\,\sigma^2 = \mathrm{tr}(\boldsymbol{A})\,\sigma^2.\end{aligned}$$

The UBRE criterion is then defined as

$$\mathcal{V}_u(\lambda) = \frac{1}{n}\left\|\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y}\right\|^2 - \sigma^2 + \frac{2\sigma^2}{n}\,\mathrm{tr}(\boldsymbol{A}), \tag{4.21}$$

where the right side depends on $\lambda$ through $\boldsymbol{A}$. As a consequence of (4.21) the minimization of $\mathcal{V}_u$ with respect to $\lambda$, minimizes the $MSE$ too.

This criterion works well if $\sigma^2$ is known. On the other hand, if $\sigma^2$ is unknown, (4.16) ensues that $\|\,\boldsymbol{y} - \boldsymbol{A}\,\boldsymbol{y}\,\|^2 = \hat{\sigma}^2(n - \mathrm{tr}(\boldsymbol{A}))$ and therefore

$$\widehat{MSE} = \frac{1}{n}\hat{\sigma}^2\mathrm{tr}(\boldsymbol{A}),$$

which turns out to be problematic. Wood (2006a) illustrates this issue by comparing a model with one parameter to one with a second parameter. Since $\mathrm{tr}(\boldsymbol{A})$ describes the degrees of freedom, the second model would have to cut $\hat{\sigma}^2$ in half to improve the $\widehat{MSE}$. As a result, models with additional parameters would seldom improve the criterion and thus get excluded. Therefore, if $\sigma^2$ is unknown, one might consider using one of the criteria from the following subsections.

### 4.3.2. Cross Validation

Since the UBRE criterion turned out to be problematic if $\sigma^2$ is unknown, in this subsection another criterion is introduced. The problem of the $MSE$ is that the true parameter $\boldsymbol{\mu}$ is unknown, therefore we now minimize the prediction error. The now presented approach to estimate the prediction error is called cross validation. For this method one observation $y_i$ is omitted in turns. After estimating the parameters without this observation, one is able to predict $y_i$. Thereby, we can get an estimate for the prediction error by calculating

$$\mathcal{V}_o(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i^{[-i]}\right)^2, \tag{4.22}$$

where $\hat{\mu}_i{}^{[-i]}$ is the fit resulting if the $i$-th observation $y_i$ is omitted.

If in (4.22) we substitute $y_i$ by $\mu_i + \varepsilon_i$, we get

$$\begin{aligned}\mathcal{V}_o(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i + \varepsilon_i - \hat{\mu}_i^{[-i]}\right)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\mu_i - \hat{\mu}_i^{[-i]}\right)^2 - 2\varepsilon_i\left(\mu_i - \hat{\mu}_i^{[-i]}\right) + \varepsilon_i^2\right).\end{aligned}$$

From $\mathbb{E}[\varepsilon_i] = 0$ and $\operatorname{var}(\varepsilon_i) = \sigma^2$ it follows that

$$\mathbb{E}[\mathcal{V}_o] = \frac{1}{n}\,\mathbb{E}\left[\sum_{i=1}^{n}\left(\mu_i - \hat{\mu}_i^{[-i]}\right)^2\right] + \sigma^2, \tag{4.23}$$

and since in the large sample limit $\hat{\mu}_i^{[-i]} \to \hat{\mu}_i$, $i = 1, \ldots, n$, it follows that the mean of the estimator of the prediction error $\mathcal{V}_o$ in (4.22) is

$$\mathbb{E}[\mathcal{V}_o] \approx MSE + \sigma^2. \tag{4.24}$$

Therefore, the minimization of $\mathcal{V}_o$ defined by (4.22) minimizes the mean squared error too. But even without this justification minimizing the prediction error turns out to be a good approach. Because if one searches a model that fits the data best, more complicated models are chosen over simpler ones. On the other hand, if we minimize the prediction error, less complicated models are chosen.

One might observe that estimating $n$ models, where in turn one observation is left out, could get computationally expensive. Fortunately, there is another way to calculate $\mathcal{V}_o$, where only one model including all $n$ observations needs to be estimated, i.e.

$$\mathcal{V}_o(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - \boldsymbol{A}_{ii})^2}. \tag{4.25}$$

The equivalence of the two representations of $\mathcal{V}_o$ can be shown by minimizing

$$\sum_{\substack{k=1\\k\neq i}}^{n}\left(y_k - \hat{\mu}_k^{[-i]}\right)^2 + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma}\,.$$

If we add $\hat{\mu}_i^{[-i]} - \hat{\mu}_i^{[-i]} = 0$ to the sum, we get

$$\sum_{k=1}^{n}\left(y_k^* - \hat{\mu}_k^{[-i]}\right)^2 + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma} \tag{4.26}$$

with $\boldsymbol{y}^* = \boldsymbol{y} - \tilde{\boldsymbol{y}}^{[i]} + \tilde{\boldsymbol{\mu}}^{[i]}$, where $\tilde{\boldsymbol{y}}^{[i]}$ and $\tilde{\boldsymbol{\mu}}^{[i]}$ are vectors whose $i$-th entries are $y_i$ and $\hat{\mu}_i^{[-i]}$ respectively while all remaining elements are zero.

If the model is estimated by minimizing (4.26), the $i$-th fitted value is $\hat{\mu}_i^{[-i]}$. Furthermore, from the definition of the influence matrix $\boldsymbol{A}$ the following equation ensues for the $i$-th fitted value:

$$\begin{aligned}\hat{\mu}_i^{[-i]} &= \boldsymbol{a}_i^T\,\boldsymbol{y}^* = \boldsymbol{a}_i^T\,\boldsymbol{y} - \boldsymbol{A}_{ii}\,y_i + \boldsymbol{A}_{ii}\,\hat{\mu}_i^{[-i]}\\ &= \hat{\mu}_i - \boldsymbol{A}_{ii}\,y_i + \boldsymbol{A}_{ii}\,\hat{\mu}_i^{[-i]},\end{aligned}$$

where $\boldsymbol{a}_i^T$ denotes the $i$-th row of $\boldsymbol{A}$.

Therefore,

$$\begin{aligned} y_i - \hat{\mu}_i^{[-i]} &= y_i - \hat{\mu}_i + \boldsymbol{A}_{ii}\, y_i - \boldsymbol{A}_{ii}\, \hat{\mu}_i^{[-i]}, \\ (1 - \boldsymbol{A}_{ii}) \left( y_i - \hat{\mu}_i^{[-i]} \right) &= y_i - \hat{\mu}_i, \\ y_i - \hat{\mu}_i^{[-i]} &= \frac{y_i - \hat{\mu}_i}{1 - \boldsymbol{A}_{ii}}. \end{aligned}$$

As a result, the two definitions of $\mathcal{V}_o$ in (4.22) and in (4.25) are equivalent.

### 4.3.3. Generalized Cross Validation

Although we found a way to estimate $\mathcal{V}_o$ without estimating $n$ models, the cross validation criterion can still get computationally expensive if more than one smooth function (see Chapter 5) is considered in the model. In addition, the lack of invariance of this criterion turns out to be a little disturbing. By the absence of invariance we mean that while minimizing

$$\| \boldsymbol{y} - \boldsymbol{X}\, \boldsymbol{\gamma} \|^2 + \lambda\, \boldsymbol{\gamma}^T\, \boldsymbol{S}\, \boldsymbol{\gamma}$$

or minimizing

$$\| \boldsymbol{Q}\, (\boldsymbol{y} - \boldsymbol{X}\, \boldsymbol{\gamma}) \|^2 + \lambda\, \boldsymbol{\gamma}^T\, \boldsymbol{S}\, \boldsymbol{\gamma}$$

result in the same parameter estimate $\hat{\boldsymbol{\gamma}}$ for any orthogonal matrix $\boldsymbol{Q}$, different cross validation scores (CV) ensue. The solution of this problem is a new criterion, namely generalized cross validation or GCV.

Since $\| \boldsymbol{y} - \hat{\boldsymbol{\mu}} \|^2$ is in this sense invariant but the elements $\boldsymbol{A}_{ii}$ in (4.25) are not, it makes sense to substitute each $\boldsymbol{A}_{ii}$ with its mean $\frac{1}{n} \operatorname{tr}(\boldsymbol{A})$. As a consequence, the resulting GCV score is invariant to rotation and given by

$$\mathcal{V}_g(\lambda) = \frac{n \| \boldsymbol{y} - \hat{\boldsymbol{\mu}} \|^2}{(n - \operatorname{tr}(\boldsymbol{A}))^2}. \tag{4.27}$$

In addition to the advantage of invariance, the GCV criterion is also computationally more efficient than the CV score. Furthermore, one can show that $(\hat{\boldsymbol{\gamma}}, \hat{\lambda})$ minimizing GCV also minimize the mean squared error $MSE$.

## 4.4. Distributional Results

Since the response vector $\boldsymbol{y}$ is assumed to be normally distributed, from (4.13) for $\lambda$ fixed the normal distribution of $\hat{\boldsymbol{\gamma}}$ follows, and $\hat{\boldsymbol{\gamma}} \sim \mathcal{N}(\mathbb{E}[\hat{\boldsymbol{\gamma}}], \boldsymbol{V}_a)$ with variance matrix

$$\begin{aligned} \boldsymbol{V}_a &= \operatorname{var}\left( \left( \boldsymbol{X}^T\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T\, \boldsymbol{y} \right) \\ &= \left( \boldsymbol{X}^T\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T \operatorname{var}(\boldsymbol{y})\, \boldsymbol{X} \left( \boldsymbol{X}^T\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1} \\ &= \sigma^2 \left( \boldsymbol{X}^T\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T\, \boldsymbol{X} \left( \boldsymbol{X}^T\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1}. \end{aligned}$$

If we consider an approach without a penalty term ($\lambda = 0$), then $\mathbb{E}[\hat{\boldsymbol{\gamma}}] = \boldsymbol{\gamma}$ and the variance matrix is $\boldsymbol{V}_a = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$. This results from the fact that without a penalty an additive model is equivalent to a linear regression model and therefore the theory in Chapter 2 including hypothesis testing and confidence and prediction intervals applies here too.

However, if a penalty is considered then the estimator $\hat{\boldsymbol{\gamma}}$ can be biased. In addition, there is no information of how the choice of $\lambda$ influences the distributional results. This is especially important for hypothesis testing. As a consequence, Wood (2006a) did simulations to analyse the context further. Overall the simulations showed good results, although it turned out that for single components confidence intervals can be unreliable.

On the other hand, Wood (2006a) argues that for each value of $\lambda$ one can find a number of basis functions $q$ for an unpenalized additive model so that the resulting fits are the same. Since we did already establish that an unpenalized additive model is equivalent to a linear regression model, the theory in Chapter 2 applies.

# 5. Generalized Additive Models

Generalized additive models or GAMs are to additive models as generalized linear models are to linear models. While in the last chapter $y$ was described by a smooth function of an explanatory variable $x$, now a link function of the mean $\mu$ is characterized by $f(x)$. Therefore, for an observation $(x, y)$ a GAM takes the form

$$g(\mu) = f(x), \tag{5.1}$$

where $\mathbb{E}[y] = \mu$, $y \sim$ Exponential family$(\theta)$ and $f(x)$ is a smooth function in the sense of a continuous function with continuous first and second derivatives of an explanatory variable $x$, and $g(\cdot)$ represents a monotone link function.

The structure of this chapter is similar as in Chapter 4, but now the theory is extended to the case of generalized additive models. Therefore, we will introduce similar smoothing techniques to get $f(\cdot)$ as in the last chapter. In addition, we again demand that $f(\cdot)$ can be written as

$$f(x) = \sum_{j=0}^{q-1} \gamma_j b_j(x) \tag{5.2}$$

with basis functions $b_j(x)$ and parameters $\gamma_j$, $j = 0, \ldots, q-1$. The definition of the basis functions is the same as in Chapter 4. Therefore, all the basis functions introduced in the last chapter can be used for generalized additive models.

While in case of additive models the consequence of (5.2) was that they are equivalent to linear regression models, it ensures that GAMs are equivalent to GLMs. This means that if we regard the section about basis functions in the last chapter, then by changing `lm` to `glm` in the R-code we get a GAM instead of an additive model. Furthermore, by specifying `family` as a member of the exponential family, we can extend the assumption of a normal distribution to any other distribution which is a member of the exponential family.

But first we want to mention that we will illustrate the theory by using the gas flow dataset obtained from the Open Grid Europe GmbH (OGE), a leading German gas transmission operator. More information about this dataset is available in Chapter 4, Chapter 6 or in Friedl et al. (2012).

Further information on generalized additive models can be found in Hastie and Tibshirani (1990), in Wood (2006a) or in Marx and Eilers (1998).

# 5.1. Penalized Splines

Similar as in Chapter 4 we want to shift the problem of choosing the number of knots to choosing a smoothing parameter $\lambda$. In that sense we start with a large number of basis functions $q$, which assures much flexibility, and then diminish the degree of flexibility by a penalty term to gain a certain degree of smoothness.

Here the method to estimate such a model, the resulting degrees of freedom and an estimator for the dispersion parameter $\phi$ are discussed. Thereafter in the next part the choice of $\lambda$ is addressed.

## 5.1.1. Model Estimation

If we are interested in a generalized additive model instead of an additive model, we proceed similar as in Chapter 3 and maximize the log-likelihood instead of minimizing the sum of squares. Regarding $n$ observations $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and assuming a model of the form

$$g(\boldsymbol{\mu}) = f(\boldsymbol{x}) = \boldsymbol{X}\,\boldsymbol{\gamma}, \tag{5.3}$$

where $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}]$ with $\boldsymbol{y} \sim$ Exponential family$(\boldsymbol{\theta})$ and the last result is due to the definition of $f(\cdot)$ in (5.2). The penalized log-likelihood function can be written as

$$l_p(\boldsymbol{y}, \boldsymbol{\gamma}) = l(\boldsymbol{y}, \boldsymbol{\gamma}) - \frac{1}{2}\lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma}, \tag{5.4}$$

where $l(\boldsymbol{y}, \boldsymbol{\gamma})$ stands for the sample log-likelihood, see Chapter 3, and the penalty term $\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma}$ is the same as in the additive model case before. To maximize (5.4), we compute

$$\begin{aligned}\frac{\partial l_p(\boldsymbol{y}, \boldsymbol{\gamma})}{\partial \gamma_j} &= \frac{\partial l(\boldsymbol{y}, \boldsymbol{\gamma})}{\partial \gamma_j} - \lambda\,\boldsymbol{s}_j^T\,\boldsymbol{\gamma} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} - \lambda\,\boldsymbol{s}_j^T\,\boldsymbol{\gamma}, \quad j = 0, \ldots, q-1,\end{aligned}$$

where $\boldsymbol{s}_j^T$ describes the $j$-th row of $\boldsymbol{S}$, $x_{ij}$ the respective element of $\boldsymbol{X}$, and the last expression is due to (3.6).

We will show that the minimization of

$$\mathcal{S}_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i(\boldsymbol{\gamma}))^2}{\operatorname{var}(y_i)} + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma}, \tag{5.5}$$

where $\operatorname{var}(y_i)$ is assumed to be fixed and $\frac{\partial \mu_i}{\partial \gamma_j} = \frac{x_{ij}}{g'(\mu_i)}$, leads to the same solution as maximizing $l_p(\boldsymbol{y}, \boldsymbol{\gamma})$, because

$$\frac{\partial \mathcal{S}_p}{\partial \gamma_j} = -2 \sum_{i=1}^{n} \frac{y_i - \mu_i}{\operatorname{var}(y_i)} \frac{x_{ij}}{g'(\mu_i)} + 2\lambda\,\boldsymbol{s}_j^T\,\boldsymbol{\gamma}\,.$$

The first part of this estimating function equals the score function under a generalized linear model and therefore the theory of Chapter 3 can be applied here. As a consequence, we can use the findings of Section 3.2 and thereby get the equivalent minimization problem

$$\mathcal{S}_p \approx \| \sqrt{\boldsymbol{W}^{(t)}} \left( \boldsymbol{z}^{(t)} - \boldsymbol{X}\,\boldsymbol{\gamma} \right) \|^2 + \lambda\, \boldsymbol{\gamma}^T\, \boldsymbol{S}\, \boldsymbol{\gamma}, \tag{5.6}$$

where $\boldsymbol{W}^{(t)}$ is a diagonal matrix and $\boldsymbol{z}^{(t)}$ a vector defined by

$$\begin{aligned} w_{ii}^{(t)} &= \frac{1}{V(\mu_i^{(t)}) g'(\mu_i^{(t)})^2}, \\ z_i^{(t)} &= g'(\mu_i^{(t)}) \left( y_i - \mu_i^{(t)} \right) + \boldsymbol{x}_i^T\, \boldsymbol{\gamma}^{(t)}. \end{aligned}$$

Therefore, the estimator $\hat{\boldsymbol{\gamma}}$ is the result of the following iteration:

- Given a current $\boldsymbol{\gamma}^{(t)}$, we calculate $\boldsymbol{\mu}^{(t)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{z}^{(t)}$ and $\boldsymbol{W}^{(t)}$.
- Minimize (5.6) with respect to $\boldsymbol{\gamma}$ to obtain $\boldsymbol{\gamma}^{(t+1)}$.
- $t \to t+1$

After the final step of the iteration at convergence we can compute the so called influence matrix $\boldsymbol{A}$, which we will use to calculate the effective degrees of freedom, the residual variance and an estimator of the dispersion parameter,

$$\boldsymbol{A} = \boldsymbol{X} (\boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{X} + \lambda\, \boldsymbol{S})^{-1}\, \boldsymbol{X}^T\, \boldsymbol{W}. \tag{5.7}$$

This definition of $\boldsymbol{A}$ is due to the penalized likelihood estimation of GAMs. Taking the first derivative of (5.4) with respect to $\boldsymbol{\gamma}$ and inserting the results of Chapter 3 leads to

$$\frac{\partial l_p(\boldsymbol{y}, \boldsymbol{\gamma})}{\partial\, \boldsymbol{\gamma}} = \frac{\partial l(\boldsymbol{y}, \boldsymbol{\gamma})}{\partial\, \boldsymbol{\gamma}} - \lambda\, \boldsymbol{S}\, \boldsymbol{\gamma} = \boldsymbol{X}^T\, \boldsymbol{D}\, \boldsymbol{W}\, (\boldsymbol{y} - \boldsymbol{\mu}) - \lambda\, \boldsymbol{S}\, \boldsymbol{\gamma}, \tag{5.8}$$

where $\boldsymbol{D}$ is a matrix with diagonal elements $d_{ii} = g'(\mu_i)$. To maximize the penalized likelihood (5.4), we at first rearrange this score function

$$\boldsymbol{X}^T\, \boldsymbol{D}\, \boldsymbol{W}\, (\boldsymbol{y} - \boldsymbol{\mu}) - \lambda\, \boldsymbol{S}\, \boldsymbol{\gamma} = \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{z} - \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{X}\, \boldsymbol{\gamma} - \lambda\, \boldsymbol{S}\, \boldsymbol{\gamma},$$

where $\boldsymbol{z} = \boldsymbol{X}\, \boldsymbol{\gamma} + \boldsymbol{D}\, (\boldsymbol{y} - \boldsymbol{\mu})$, and then set the last term above equal to zero

$$\boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{z} - \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{X}\, \boldsymbol{\gamma} - \lambda\, \boldsymbol{S}\, \boldsymbol{\gamma} = \boldsymbol{0},$$

and thereby derive

$$\boldsymbol{\gamma}^{(t+1)} = \left( \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{z}, \tag{5.9}$$

where the right side is evaluated in $\boldsymbol{\gamma}^{(t)}$ and $t$ represents the current iteration number. At convergence the estimator $\hat{\boldsymbol{\gamma}}$ results from $\boldsymbol{\gamma}^{(t+1)}$.

Furthermore,

$$\begin{aligned} \boldsymbol{\eta} = \boldsymbol{X}\, \boldsymbol{\gamma} &= \boldsymbol{X} \left( \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{X} + \lambda\, \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T\, \boldsymbol{W}\, \boldsymbol{z} \\ &= \boldsymbol{A}\, \boldsymbol{z}, \end{aligned}$$

where $\boldsymbol{A}$ is defined as in (5.7).

### 5.1.2. Effective Degrees of Freedom and Dispersion Parameter

Similarly as in Chapter 4 now the question how many degrees of freedom a GAM has is addressed. If $\lambda = 0$ the degrees of freedom are composed of the number of parameters. On the other hand, if $\lambda$ is large, the model gets very inflexible and has less degrees of freedom. In the following the matrices $\boldsymbol{W}$ and $\boldsymbol{A}$ are the matrices resulting at convergence from the iterative weighed least squares approach described earlier. Similarly $\hat{\boldsymbol{z}}$ and $\hat{\boldsymbol{\gamma}}$ result from $\boldsymbol{\gamma}^{(t)}$ and $\boldsymbol{z}^{(t)}$ at convergence.

Again the effective degrees of freedom are defined as the trace of the hat matrix or in this case the influence matrix $\boldsymbol{A}$, i.e.

$$p = \mathrm{tr}(\boldsymbol{A}).$$

If we are interested in the degrees of freedom of the parameter $\boldsymbol{\gamma}$, then in the generalized additive model case we define

$$\boldsymbol{P} = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}.$$

Thus, $\hat{\boldsymbol{\gamma}} = \boldsymbol{P}\,\hat{\boldsymbol{z}}$ and $\boldsymbol{A} = \boldsymbol{X}\,\boldsymbol{P}$, and the trace of $\boldsymbol{A}$ can be written as

$$\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{X}\,\boldsymbol{P}) = \sum_{i=1}^{n} (\boldsymbol{P}\,\boldsymbol{X})_{ii}.$$

In other words, the sum of the diagonal elements of the matrix

$$\boldsymbol{P}\,\boldsymbol{X} = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$$

describes the effective degrees of freedom.

Next an estimator of the dispersion parameter is introduced. Similar as in Chapter 4 the residual degrees of freedom of a GLM, $n-p$, are substituted by the residual effective degrees of freedom, namely $n - \mathrm{tr}(\boldsymbol{A})$, and as a result the Pearson estimator for the dispersion parameter in a generalized additive model is given by

$$\hat{\phi} = \frac{1}{n - \mathrm{tr}(\boldsymbol{A})} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

### 5.1.3. Penalized Splines in R

If one wants to estimate P-splines with a difference penalty as described in Chapter 4 or cubic splines with a penalty on the second derivative in R, the function `gam` of the package `mgcv` turns out to be useful. Below we describe how a generalized additive model can be fitted in R based on the example of the gas flow data.

### P-splines

We consider an additive or a generalized additive model with B-splines as basis functions. In addition, a difference penalty as defined in (4.18) is used. The estimation of such a P-spline for the gas flow data can be achieved in R with the following R-code:

```
library(mgcv)
gam(max.flow~s(temp,bs="ps",m=c(2,2)))
```

After loading the package `mgcv`, we can set up and estimate the generalized additive model with the function `gam`. The maximal gas flow is estimated as a smooth function of the temperature. Therefore, the model formula is

```
max.flow ~ s(temp)
```

where `s()` indicates the smooth function. The rest of the arguments in the R-code above define which basis functions should be used and the degree of the spline and the penalty, respectively. In our example,

```
bs="ps"
```

defines that P-splines with a penalty on the parameter differences are used like they were introduced by Eilers and Marx (1996). While in

```
m=c(2,2)
```

the first number in the argument defines the degree of the spline, the second number corresponds to the order of the differences in the penalty. In our example `m[1]=2`, which corresponds to a cubic spline. Furthermore `m[2]=2`, meaning that second order differences are used for the penalty. One might notice inconsistencies to the previous definition of the degree of a spline regarding the command `splineDesign` in R. Therefore, we recommend to look in the various help pages of R to check the current definition of the degree of a spline for the used command.

The choice of cubic B-splines and second order differences is the default setting for P-splines. If only one value is determined like `m=2`, the function `gam` assumes that the degree of the spline and the difference order of the penalty are the same, in this case two.

By default, the function `gam` automatically chooses the dimension of the basis $q$ as the maximum of 10 and `m[1]+1`. If one prefers a different number of basis functions for the smooth function this can be arranged by specifying `k` in

```
gam(max.flow~s(temp,bs="ps",m=c(2,2),k=5))
```

As a consequence of the number of basis functions $q$, a total number of `q+m[1]+2` knots are used to construct the smooth fit. Therefore, if one wants to use specific knots for the fit, there need to be `q+m[1]+2` of them and they can be set by

```
gam(max.flow~s(temp,bs="ps",m=2,k=5),knots=list(temp=knots(-3:5)))
```

Since we chose a number of `k=5` basis functions and cubic splines `m[1]=2`, we need to specify 9 knots to set up the B-splines. The knots in the R-code above are in a list. This is required by `gam`, so that if the model contains more smooth functions, each smooth function has its own specified set of knots.

Further specifications used in the function `gam` are for example the used dataset `data=...` or the exponential family `family=....` In the former R-code the default setting for the exponential family is `gaussian` with an identity link, which is equivalent to an additive model. Everyone who has already estimated GLMs in R will find that the implemented exponential families for `gam` are the same as in `glm`. In the course of this chapter and in Chapter 6 we use a gamma distribution with a log link.

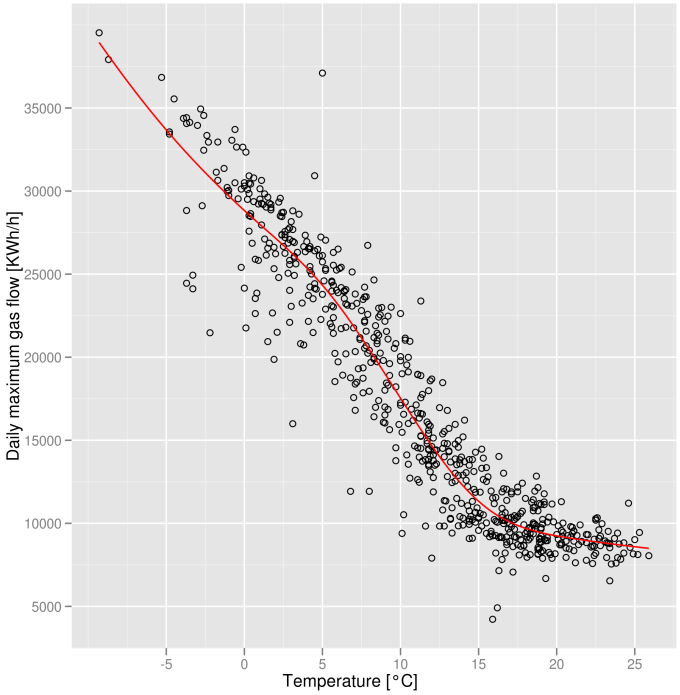


Figure 5.1.: Cubic P-spline with difference order $k = 2$ estimated by `gam` with $q = 10$ and $\lambda = 0.1038$.

In Figure 5.1 the fit resulting from the R-code

```
mod1<-gam(max.flow~s(temp,bs="ps",m=c(2,2)),family=Gamma(link=log))
```

is shown. One can observe that the fit describes the overall trend of the data well while it is not overly wiggly. Therefore, the estimated smoothing parameter $\lambda$ seems to be well-chosen. In this case the smoothing parameter $\lambda$ is estimated to be 0.1038 and can be derived from the model as

```
mod1$sp
```

where `sp` stands for smoothing parameter. The theory how to estimate $\lambda$ will be discussed later in 5.2.

### Penalized Cubic Splines

If one wants to use cubic splines with a penalty on the second derivative like in 5.1.1, the R-code can be adapted to

```
gam(max.flow~s(temp,bs="cr"),family=Gamma(link=log))
```

One might observe that instead of `bs="ps"` the basis is now chosen to be `bs="cr"`, which results in a cubic spline with a penalty equal to the penalty in (4.11).

By specifying `k`, it is again possible to influence the number of basis functions that are considered for the fit. In contrast to P-splines in this case cubic splines are selected automatically and it is therefore not possible to change the degree of the spline. In addition, the penalty is always the integral of the squared second derivative and can only be changed by choosing a different basis.

The fit resulting from this choice of basis functions and penalty is presented in Figure 5.2. Similar to P-splines the fit describes the data well without showing over- or undersmoothing, which is due to the optimal choice of $\lambda$. The smoothing parameter $\lambda$ determined by `gam` is now 2.133.

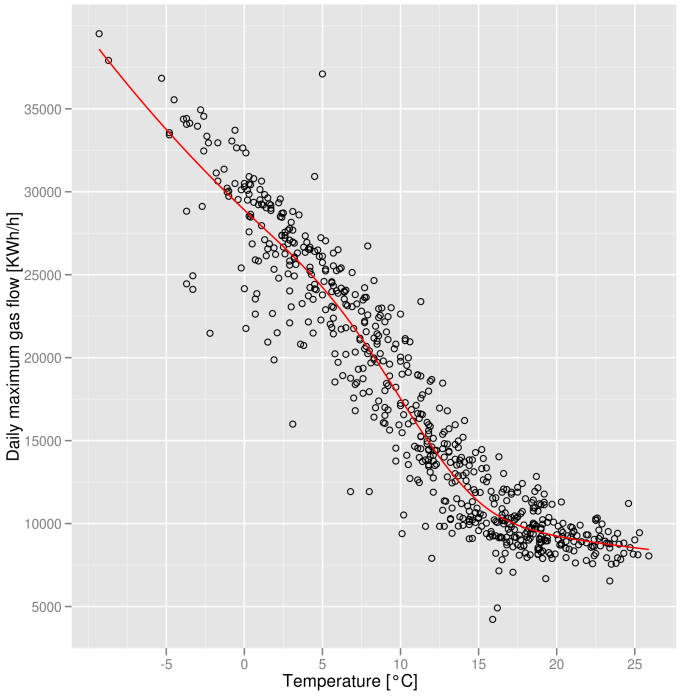


Figure 5.2.: Cubic spline with a penalty on the squared second derivative estimated by `gam` with $q = 10$ and $\lambda = 2.133$.

Next, a comparison between the two fits is regarded. In Figure 5.3 the P-spline fit is represented by a blue line while the red line stands for the penalized cubic spline. One might notice that they are almost indistinguishable except at the left and right margin,

where less observations are found. Therefore, the reasoning in 4.2.3 that there is no big difference between a penalty on the second derivative and a penalty on the second differences of the parameters seems coherent.

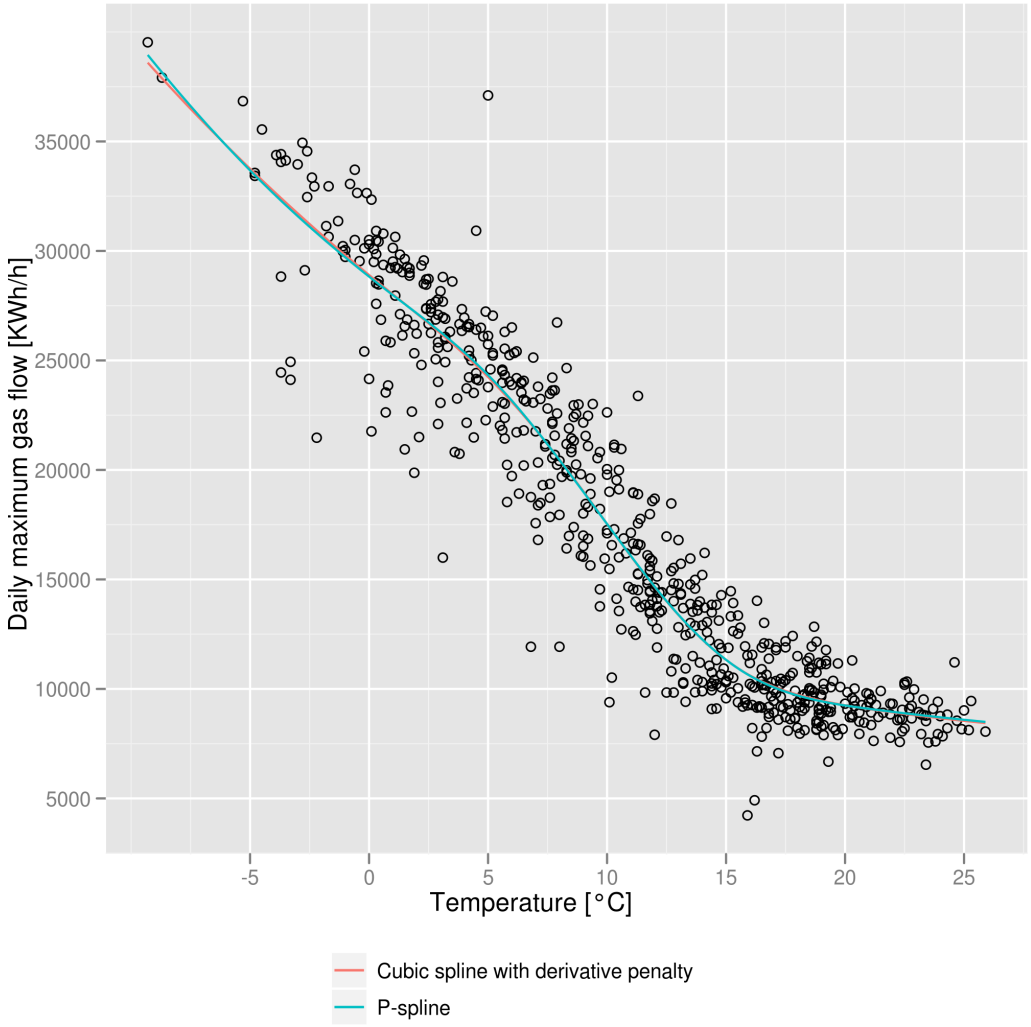


Figure 5.3.: Comparison of a cubic P-spline with difference order $k = 2$ and a penalized cubic spline with a penalty on the squared second derivative, where both are fitted using `gam` and $q = 10$.

## 5.2. Smoothing Parameter $\lambda$

In Section 5.1 we could shift the problem of choosing the number of knots to the problem of choosing the smoothing parameter $\lambda$. This section introduces criteria as to how $\lambda$ should be chosen.

### 5.2.1. Unbiased Risk Estimator

The first criterion for the choice of $\lambda$ which we discuss is the unbiased risk estimator or UBRE. Similar as in Chapter 4 the idea is to choose $\lambda$ so that $\hat{\boldsymbol{\mu}}$ is close to the true parameter $\boldsymbol{\mu}$. By regarding the mean squared error $MSE$, we did deduce the UBRE score for the additive model case in Chapter 4. In the following, the UBRE criterion is defined for a generalized additive model similarly as in the additive model case.

While in Chapter 4 the UBRE score for an additive model is defined as (see (4.21))

$$\mathcal{V}_u(\lambda) = \frac{1}{n} \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 - \sigma^2 + \frac{2\sigma^2}{n} \operatorname{tr}(\boldsymbol{A}),$$

with $\hat{\boldsymbol{\mu}} = \boldsymbol{A}\,\boldsymbol{y}$ for a GAM the UBRE criterion can be written as

$$\mathcal{V}_u^d(\lambda) = \frac{1}{n} D(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) - \sigma^2 + \frac{2\sigma^2}{n} \operatorname{tr}(\boldsymbol{A}),$$

where $D(\boldsymbol{y}, \hat{\boldsymbol{\gamma}})$ stands for the deviance as defined in Chapter 3. In the following we will also use the penalized deviance $D_p(\boldsymbol{y}, \hat{\boldsymbol{\gamma}})$, which is similarly defined as the deviance but the log-likelihood $l(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is replaced by the penalized log-likelihood $l_p(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$, i.e.

$$D_p(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) = -2\left(l_p(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) - l(\boldsymbol{y}, \boldsymbol{y})\right). \tag{5.10}$$

The difference between $\mathcal{V}_u^d(\lambda)$ and $\mathcal{V}_u(\lambda)$ is a direct result from the difference in estimation between an additive and a generalized additive model. Since in case of a GAM a log-likelihood approach is pursued instead of a least squares method as in the additive model case, the definition of the UBRE criterion changes likewise. In addition, the maximization of the penalized log-likelihood $l_p(\boldsymbol{y}, \boldsymbol{\mu})$ is equivalent to the minimization of the penalized deviance

$$\begin{aligned} D_p(\boldsymbol{y}, \boldsymbol{\gamma}) &= -2\left(l_p(\boldsymbol{y}, \boldsymbol{\mu}) - l(\boldsymbol{y}, \boldsymbol{y})\right) = -2\left(l(\boldsymbol{y}, \boldsymbol{\mu}) - l(\boldsymbol{y}, \boldsymbol{y})\right) + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma} \\ &= D(\boldsymbol{y}, \boldsymbol{\gamma}) + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma}, \end{aligned}$$

which is a result of the definition of the deviance above.

As a consequence of the change from minimizing the sum of squares to minimizing the deviance, the first term of $\mathcal{V}_u(\lambda)$ can be substituted by its analogon - the deviance. Therefore, the definition of $\mathcal{V}_u^d$ above follows.

One might also notice that $\mathcal{V}_u^d$ is a linear transformation of the Akaike Information Criterion or $AIC$, which is

$$\begin{aligned} AIC &= -2\,l_p(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) + k \operatorname{tr}(\boldsymbol{A}) \\ &= -2\left(l_p(\boldsymbol{y}, \hat{\boldsymbol{\gamma}}) - l(\boldsymbol{y}, \boldsymbol{y})\right) - 2l(\boldsymbol{y}, \boldsymbol{y}) + k \operatorname{tr}(\boldsymbol{A}), \end{aligned}$$

where $k > 0$. Usually $k$ is taken to be two, which is for example the default setting in R. As the $AIC$ is often used to choose between different models, its equivalence to the UBRE criterion is welcome.

Another similar representation of this criterion in case of a GAM is given by

$$\mathcal{V}_u^w(\lambda) = \frac{1}{n}\|\sqrt{\boldsymbol{W}}\,(\boldsymbol{z} - \boldsymbol{X}\,\hat{\boldsymbol{\gamma}})\,\|^2 - \sigma^2 + \frac{2\sigma^2}{n}\,\mathrm{tr}(\boldsymbol{A}).$$

This definition is due to the approximation of a penalized log-likelihood as an iterative reweighed least squares approach, see (5.6). In 5.1.1 we showed that maximizing the penalized log-likelihood leads approximately to the same result as minimizing the iterative reweighed least squares approach. As a result, the definition of $\mathcal{V}_u^w$ above ensues.

As in the additive model case the UBRE criterion performs well if the variance $\sigma^2$ is known but gets problematic if it is unknown. Therefore, in the next part the generalized cross validation score (GCV) is introduced.

### 5.2.2. Generalized Cross Validation

Similarly as for the UBRE criterion the GCV criterion can be derived from the analogue in the additive model case, i.e.

$$\mathcal{V}_g(\lambda) = \frac{n\,\|\,\boldsymbol{y} - \hat{\boldsymbol{\mu}}\,\|^2}{(n - \mathrm{tr}(\boldsymbol{A}))^2}.$$

But instead of minimizing the penalized sum of squares, here we minimize

$$D(\boldsymbol{y}, \boldsymbol{\gamma}) + \lambda\,\boldsymbol{\gamma}^T\,\boldsymbol{S}\,\boldsymbol{\gamma},$$

where $D(\boldsymbol{y}, \boldsymbol{\gamma})$ describes the deviance similar to the definition in Chapter 3 (see (3.12)). One might notice that according to the definition of the penalized deviance in (5.10) there is no difference whether the penalized log-likelihood is maximized or the penalized deviance minimized with respect to $\boldsymbol{\gamma}$.

Similar as in case of the UBRE criterion the deviance or the iterative re-weighed least squares method substitute their analogue in the additive case. Thereby we receive the GCV score for a GAM. In accordance with the definition of the GCV score $\mathcal{V}_g(\lambda)$ above, the score is now defined by

$$\mathcal{V}_g^d(\lambda) = \frac{n\,D\,(\boldsymbol{y}, \hat{\boldsymbol{\gamma}})}{(n - \mathrm{tr}(\boldsymbol{A}))^2}. \tag{5.11}$$

In the same way the approximation of the maximization problem of the penalized log-likelihood as an iterative reweighed least squares approach leads to the definition of

$$\mathcal{V}_g^w(\lambda) = \frac{n\,\|\sqrt{\boldsymbol{W}}\,(\boldsymbol{z} - \boldsymbol{X}\,\hat{\boldsymbol{\gamma}})\,\|^2}{(n - \mathrm{tr}(\boldsymbol{A}))^2},$$

where $\mathcal{V}_g^w$ is only locally valid since $\boldsymbol{W}$ and $\boldsymbol{z}$ depend on the current $\lambda$.

More information on GCV can be found in Hastie and Tibshirani (1990) or in Wood (2006a).

### 5.2.3. Realisation in R

In R the parameter `method`, which can be specified in `gam`, selects the criterion which shall be used for the smoothing parameter estimation. Among others the UBRE and GCV criteria are also implemented. By default GCV is used in case of an unknown dispersion parameter $\phi$, while the UBRE criterion is used if the dispersion parameter is known.

## 5.3. Distributional Results

In this section we derive a variance matrix for the estimator $\hat{\boldsymbol{\gamma}}$ and some distributional results. In case of the variance matrix of $\hat{\boldsymbol{\gamma}}$ we use a similar approach as for GLMs in Section 3.3.

Performing a Taylor expansion of the score function in the true parameter $\boldsymbol{\gamma}$ yields to

$$\mathbf{0} = \left.\frac{\partial l_p(\boldsymbol{y},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right|_{\hat{\boldsymbol{\gamma}}} \approx \left.\frac{\partial l_p(\boldsymbol{y},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}} + \left.\frac{\partial^2 l_p(\boldsymbol{y},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}\,\partial \boldsymbol{\gamma}^T}\right|_{\boldsymbol{\gamma}} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\,.$$

We did already establish in equation (5.8) that the first derivative of the penalized log-likelihood takes the form

$$\frac{\partial l_p(\boldsymbol{y},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} = \boldsymbol{X}^T \boldsymbol{D}\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}) - \lambda \boldsymbol{S}\boldsymbol{\gamma}\,.$$

Since the first term of the score function above is the same as in case of a GLM, we can use the result of Chapter 3 for the first term of the second derivative, i.e.

$$-\frac{\partial^2 l_p(\boldsymbol{y},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}\,\partial \boldsymbol{\gamma}^T} = \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + \lambda \boldsymbol{S}\,.$$

Therefore, the approximation above can be written as

$$\mathbf{0} \approx \boldsymbol{X}^T \boldsymbol{D}\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}) - \lambda \boldsymbol{S}\boldsymbol{\gamma} - \left(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + \lambda \boldsymbol{S}\right)(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\,,$$

resulting in

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \approx \left(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1}\left(\boldsymbol{X}^T \boldsymbol{D}\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}) - \lambda \boldsymbol{S}\boldsymbol{\gamma}\right).$$

The bias of $\hat{\boldsymbol{\gamma}}$ resulting from this approximation is therefore

$$\mathbb{E}\left[\hat{\boldsymbol{\gamma}}\right] - \boldsymbol{\gamma} \approx -\left(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \lambda \boldsymbol{S}\boldsymbol{\gamma}\,.$$

For $\lambda = 0$ it follows that $\mathbb{E}\left[\hat{\boldsymbol{\gamma}}\right] = \boldsymbol{\gamma}$, while for $\lambda \neq 0$ this is not necessarily true.

In addition to the bias we can also derive a variance matrix for $\hat{\boldsymbol{\gamma}}$, i.e.

$$\operatorname{var}(\hat{\boldsymbol{\gamma}}) = \boldsymbol{V}_e = \left(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1} \boldsymbol{X}^T \boldsymbol{D}\boldsymbol{W} \operatorname{var}(\boldsymbol{y})\, \boldsymbol{W}\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} + \lambda \boldsymbol{S}\right)^{-1}.$$

Since

$$\operatorname{var}(\boldsymbol{y}) = \phi V(\boldsymbol{\mu}) = \phi \, \boldsymbol{D}^{-1} \boldsymbol{W}^{-1} \boldsymbol{D}^{-1},$$

it follows that

$$\boldsymbol{V}_e = \phi \left( \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \lambda \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \left( \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \lambda \boldsymbol{S} \right)^{-1}. \tag{5.12}$$

Next we want to derive some distributional results for $\hat{\boldsymbol{\gamma}}$. Since the estimator

$$\hat{\boldsymbol{\gamma}} = \left( \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} + \lambda \boldsymbol{S} \right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$$

in (5.9) depends on $\boldsymbol{z}$, where $\boldsymbol{z} = \boldsymbol{X} \boldsymbol{\gamma} + \boldsymbol{D}(\boldsymbol{y} - \boldsymbol{\mu})$, a distributional result for $\boldsymbol{z}$ is needed. Later in Theorem 2 we will show that approximately

$$\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z} \xrightarrow{d} \mathcal{N} \left( \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\gamma}, \phi \, \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \right).$$

For this result the following equivalence and Theorem 2 are used: If $\mathbf{v} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$ is multivariate normally distributed, then $\mathbf{c}^T \mathbf{v}$ is also multivariate normally distributed for any vector of constants $\mathbf{c} \neq \mathbf{0}$ and the other way round.

**Theorem 2.** *Under certain regularity conditions the random variable* $\mathbf{c}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$*, for any vector* $\mathbf{c} \neq \mathbf{0}$*, converges in distribution towards a multivariate normal distribution.*

The proof can be found in Appendix B or in Wood (2006a).

The large sample multivariate normality of $\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$ and (5.9) ensue that approximately

$$\hat{\boldsymbol{\gamma}} \xrightarrow{d} \mathcal{N} \left( \mathbb{E}\left[ \hat{\boldsymbol{\gamma}} \right], \boldsymbol{V}_e \right), \tag{5.13}$$

where usually $\mathbb{E}\left[ \hat{\boldsymbol{\gamma}} \right] \neq \boldsymbol{\gamma}$, as shown before.

### 5.3.1. Confidence and Prediction Intervals

The variance matrix of $\hat{\boldsymbol{\mu}}$ can be derived from the variance matrix $\boldsymbol{V}_e$ of $\hat{\boldsymbol{\gamma}}$ using the so called delta method, see proof of Theorem 3. The following theorem is available in Casella and Berger (2002).

**Theorem 3.** *If a series of random variables* $\hat{\boldsymbol{\gamma}}_n$ *with variance matrix* $\boldsymbol{\Sigma}$ *satisfies*

$$\sqrt{n} \left( \hat{\boldsymbol{\gamma}}_n - \mathbb{E}\left[ \hat{\boldsymbol{\gamma}}_n \right] \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

*then for any real-valued function* $h(\cdot)$ *with* $\nabla h(\mathbb{E}\left[ \hat{\boldsymbol{\gamma}}_n \right]) \neq \mathbf{0}$

$$\sqrt{n} \left( h\left( \hat{\boldsymbol{\gamma}}_n \right) - h\left( \mathbb{E}\left[ \hat{\boldsymbol{\gamma}}_n \right] \right) \right) \xrightarrow{d} \mathcal{N} \left( \mathbf{0}, \nabla h \left( \mathbb{E}\left[ \hat{\boldsymbol{\gamma}}_n \right] \right)^T \boldsymbol{\Sigma} \, \nabla h \left( \mathbb{E}\left[ \hat{\boldsymbol{\gamma}}_n \right] \right) \right).$$

*Proof.* This theorem can be proven by considering a Taylor approximation of $h(\cdot)$ in $\mathbb{E}[\hat{\boldsymbol{\gamma}}_n]$

$$h(\hat{\boldsymbol{\gamma}}_n) \approx h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n]) + \nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])^T (\hat{\boldsymbol{\gamma}}_n - \mathbb{E}[\hat{\boldsymbol{\gamma}}_n]).$$

Rearranging the Taylor approximation above and multiplying it with $\sqrt{n}$ leads to

$$\sqrt{n}\,(h(\hat{\boldsymbol{\gamma}}_n) - h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])) \approx \nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])^T \sqrt{n}\,(\hat{\boldsymbol{\gamma}}_n - \mathbb{E}[\hat{\boldsymbol{\gamma}}_n]).$$

While $\nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])$ on the right side is constant, $\sqrt{n}\,(\hat{\boldsymbol{\gamma}}_n - \mathbb{E}[\hat{\boldsymbol{\gamma}}_n])$ is assumed to tend to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, and therefore the convergence of $\sqrt{n}\,(h(\hat{\boldsymbol{\gamma}}_n) - h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n]))$ follows, where the variance matrix is given by

$$\begin{aligned}
\sqrt{n}\,\mathrm{var}\,(h(\hat{\boldsymbol{\gamma}}_n)) &\approx \sqrt{n}\,\mathrm{var}\left(\nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])^T (\hat{\boldsymbol{\gamma}}_n - \mathbb{E}[\hat{\boldsymbol{\gamma}}_n])\right) \\
&\approx \nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])^T \sqrt{n}\,\mathrm{var}\,(\hat{\boldsymbol{\gamma}}_n)\,\nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n]) \\
&\approx \nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n])^T \boldsymbol{\Sigma}\,\nabla h(\mathbb{E}[\hat{\boldsymbol{\gamma}}_n]).
\end{aligned}$$

□

### Confidence Interval

Together Theorem 2 and Theorem 3 yield to an approximate confidence interval. From the approximate normal distribution of $\hat{\boldsymbol{\gamma}}$ in Theorem 2, the variance matrix $\boldsymbol{V}_e$ as defined in (5.12) and Theorem 3 it follows that

$$\sqrt{n}\,(\hat{\boldsymbol{\mu}} - \mathbb{E}[\hat{\boldsymbol{\mu}}]) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, n\left(\nabla g^{-1}(\mathbb{E}[\hat{\boldsymbol{\gamma}}])\right)^T \boldsymbol{V}_e\,\nabla g^{-1}(\mathbb{E}[\hat{\boldsymbol{\gamma}}])\right), \tag{5.14}$$

since

$$\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{X}\,\hat{\boldsymbol{\gamma}}),$$

where $g^{-1}(\cdot)$ describes the inverse link function. As a result of (5.14) we know that approximately

$$\left(\left(\nabla g^{-1}(\mathbb{E}[\hat{\boldsymbol{\gamma}}])\right)^T \boldsymbol{V}_e\,\nabla g^{-1}(\mathbb{E}[\hat{\boldsymbol{\gamma}}])\right)^{-\frac{1}{2}} (\hat{\boldsymbol{\mu}} - \mathbb{E}[\hat{\boldsymbol{\mu}}]) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

and therefore the $(1-\alpha)$ confidence interval for $\mathbb{E}[\hat{\mu}_i]$ if $\phi$ is known is given by

$$\left(\hat{\mu}_i - \sqrt{v_i} z_{1-\alpha/2}, \hat{\mu}_i + \sqrt{v_i} z_{1-\alpha/2}\right),$$

where $v_i$ is the $i$-th diagonal element of $\left(\nabla g^{-1}(\mathbb{E}[\hat{\boldsymbol{\gamma}}])\right)^T \boldsymbol{V}_e\,\nabla g^{-1}(\mathbb{E}[\hat{\boldsymbol{\gamma}}])$.

However, if the dispersion parameter $\phi$ in $\boldsymbol{V}_e$ is unknown (see (5.12)) then replacing $\phi$ by $\hat{\phi}$ results in a confidence interval of the form

$$\left(\hat{\mu}_i - \sqrt{v_i} t_{n-p,1-\alpha/2}, \hat{\mu}_i + \sqrt{v_i} t_{n-p,1-\alpha/2}\right),$$

where $v_i$ is the $i$-th diagonal element of $\left(\nabla g^{-1}\left(\mathbb{E}\left[\hat{\boldsymbol{\gamma}}\right]\right)\right)^T \hat{\boldsymbol{V}}_e \nabla g^{-1}\left(\mathbb{E}\left[\hat{\boldsymbol{\gamma}}\right]\right)$ and $\hat{\boldsymbol{V}}_e$ results from $\boldsymbol{V}_e$ by replacing $\phi$ with $\hat{\phi}$.

Another possibility to calculate $\operatorname{var}\left(\hat{\boldsymbol{\mu}}\right)$ is by considering $\operatorname{var}\left(\hat{\boldsymbol{\mu}}\right) = \mathbb{E}\left[\hat{\boldsymbol{\mu}}^2\right] - \mathbb{E}\left[\hat{\boldsymbol{\mu}}\right]^2$. Since the two expectations are basically integrals, we can compute them by calculating

$$\mathbb{E}\left[\hat{\boldsymbol{\mu}}\right] = \mathbb{E}\left[g^{-1}\left(\boldsymbol{X}\,\hat{\boldsymbol{\gamma}}\right)\right] = \int_{-\infty}^{\infty} g^{-1}\left(\boldsymbol{X}\,\hat{\boldsymbol{\gamma}}\right) f\left(\hat{\boldsymbol{\gamma}}\right) \mathrm{d}\,\hat{\boldsymbol{\gamma}},$$
$$\mathbb{E}\left[\hat{\boldsymbol{\mu}}^2\right] = \mathbb{E}\left[\left(g^{-1}\left(\boldsymbol{X}\,\hat{\boldsymbol{\gamma}}\right)\right)^2\right] = \int_{-\infty}^{\infty} \left(g^{-1}\left(\boldsymbol{X}\,\hat{\boldsymbol{\gamma}}\right)\right)^2 f\left(\hat{\boldsymbol{\gamma}}\right) \mathrm{d}\,\hat{\boldsymbol{\gamma}},$$

where $f\left(\hat{\boldsymbol{\gamma}}\right)$ is the density function of $\hat{\boldsymbol{\gamma}}$. Since we did show in Theorem 2 that $\hat{\boldsymbol{\gamma}}$ is in the large sample limit approximately normally distributed, the means above can be approximated using the density function of the normal distribution for $f\left(\hat{\boldsymbol{\gamma}}\right)$.

During this section we discussed distributional results of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}$ under the assumption that $\lambda$ is fixed. However, in reality $\lambda$ is estimated by the data through various criteria, see Section 5.2. Since the effect the choice of $\lambda$ has on the distributional results is not clear, Wood (2006a) did simulations to further analyse the context. While Wood (2006a) concludes that the distributional results cover the results from the simulations fair enough for the whole model, single components exposed quite unreliable results. To receive better results Wood (2006a) includes the uncertainty that is connected with $\lambda$ in a Bayesian approach. Further information on confidence intervals of GAMs can be found in Wood (2006b) or in Marra and Wood (2012).

### Prediction Interval

We now want to derive a prediction interval for a new response $y^*$ to be observed at $x = x^*$. In this context we consider the new values of the explanatory variables and evaluate the basis functions $b_i(\cdot)$ in them, and thereby we get $\boldsymbol{x}^* = \left(b_0^*, \ldots, b_{q-1}^*\right)^T$. As a result from our previous findings and the independence of the new observation $y^*$ from the previous observations and the estimation process, we know that

$$\operatorname{var}\left(y^* - \hat{\mu}^*\right) = \phi V(\mu^*) + \left(\left(g^{-1}\left(\boldsymbol{x}^{*T}\,\mathbb{E}\left[\hat{\boldsymbol{\gamma}}\right]\right)\right)'\right)^2 \boldsymbol{V}_e^*,$$

where $\boldsymbol{V}_e^*$ derives from $\boldsymbol{V}_e$ by substituting $\boldsymbol{X}$ by $\boldsymbol{x}^*$. Since the distribution of $y$ is assumed to be a member of the exponential family, $\operatorname{var}(y^*) = \phi V(\mu^*)$.

If $y$ is from the normal distribution or at least symmetrically distributed around the mean

$$\left(\mu^* - z_{1-\alpha/2}\left(\phi V(\mu^*) + \operatorname{var}(\hat{\mu}^*)\right)^{\frac{1}{2}}, \mu^* + z_{1-\alpha/2}\left(\phi V(\mu^*) + \operatorname{var}(\hat{\mu}^*)\right)^{\frac{1}{2}}\right)$$

is considered for a prediction interval if $\phi$ is known with

$$\operatorname{var}(\hat{\mu}^*) = \left(\left(g^{-1}\left(\boldsymbol{x}^{*T}\,\mathbb{E}\left[\hat{\boldsymbol{\gamma}}\right]\right)\right)'\right)^2 \boldsymbol{V}_e^*.$$

If however $\phi$ is unknown, we consider a $t$-distribution instead of a normal distribution for the prediction interval, i.e.

$$\left(\mu^* - t_{n-p,1-\alpha/2}\left(\phi V(\mu^*) + \text{var}(\hat{\mu}^*)\right)^{\frac{1}{2}}, \mu^* + t_{n-p,1-\alpha/2}\left(\phi V(\mu^*) + \text{var}(\hat{\mu}^*)\right)^{\frac{1}{2}}\right).$$

Since the response variable which we will analyse in Chapter 6 consists of daily maxima, we will consider a gamma distribution there. A gamma distribution also allows the data to be asymmetrically shaped. As a result the prediction intervals introduced above might not be appropriate. In this case we will use the following approach: Using the estimated mean and variance we will calculate $(1-\alpha)$-quantiles of the gamma distribution. Thereby we get approximate prediction intervals, which can be compared to prediction intervals based on the symmetric assumption above. However, we want to note that the prediction intervals resulting from these quantiles do not consider the variability of the estimate of the mean, which can result in prediction intervals that are too narrow. For the application of this strategy we refer to Chapter 6.

### 5.3.2. Hypothesis Testing

Since GAMs are similar to GLMs, it makes sense to use the likelihood ratio statistic as test statistic for nested models. In context of a GAM the likelihood ratio statistic takes the form

$$\frac{1}{\phi}\left(D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)\right),$$

where $\hat{\boldsymbol{\mu}}_0$ and $\hat{\boldsymbol{\mu}}_1$ denote the fitted values under the null and alternative hypothesis, respectively, see Section 3.6. In addition, in the definition of the deviance in case of a GAM the log-likelihood $l(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ is replaced by the penalized log-likelihood $l_p(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ if a penalized approach is considered for the parameter estimation.

The problem of this approach is that in case of P-splines the distribution of the likelihood ratio is unknown. Wood (2006a) suggests that under the assumption of a known smoothing parameter $\lambda$, the likelihood ratio should approximately be distributed as in case of a GLM, and therefore

$$\frac{1}{\phi}\left(D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)\right) \sim \chi^2_{p_0-p_1}, \tag{5.15}$$

where $p_0$ and $p_1$ describe the effective degrees of freedom under the null and alternative hypothesis, respectively.

Wood (2006a) reinforces this assumption by illustrating that every estimate resulting from a P-splines approach can also be generated by usual regression splines (without a penalty), if the same degrees of freedom are ensured. Since a GAM with regression splines is the same as a GLM, the theory in Chapter 3 applies and (5.15) is approximately true. Finally, Wood (2006a) concludes that since P-splines and regression splines with the same degrees of freedom lead to a nearly equal fit, the deviance depending on the fit should be nearly the same too. As a result, the test statistic should be similar in both cases and (5.15) might hold approximately true for P-splines.

## 5.4. Extrapolation

Since in the practical example in Chapter 6 we will be interested in the behaviour of the mean outside the observed range of the explanatory variable, this section deals with the topic of extrapolation. In Currie and Durban (2002) a GAM using P-splines is estimated for mortality rates. Later in Currie, Durban, and Eilers (2003) they address the problem of forecasting future mortality rates. In this part we introduce a similar approach for extrapolation as they used for forecasting.

If we want to extrapolate P-splines or GAMs on a data range $[a^*, b^*]$ adjoining the original data range $[a, b]$, we add basis functions $b_l(x)$, $l = 0, \ldots, q^* - 1$ on $[a^*, b^*]$ and estimate the function $f^*(\cdot)$ on $[a^*, b^*]$, which is an extrapolation of $f(x)$ on $[a, b]$, as

$$f^*(x) = \sum_{l=0}^{q^*-1} \gamma_l^* b_l(x),$$

where the basis functions $b_l(x)$ are chosen similarly as in the model $f(x) = \sum_{j=0}^{q-1} \gamma_j b_j(x)$. The only remaining problem lies in the fact that for the new basis functions new parameters need to be estimated.

To solve the problem of how to estimate the new parameters $\boldsymbol{\gamma}^* = (\gamma_0^*, \ldots, \gamma_{q^*-1}^*)^T$, we remember that the penalized log-likelihood was used to estimate the model, see (5.4). Since no additional observations are added, the first part of the penalized log-likelihood, namely the log-likelihood, stays the same. On the other hand, the penalty term can be extended to the new parameters. By keeping the already estimated parameters fixed and minimizing the penalty with respect to the new parameters, we ensure that the new parameters are a continuous extension of the already estimated model parameters.

For example, if we are interested in a continuation of the smooth function on the left side of $[a, b]$, therefore assuming that $b^* = a$, then $\boldsymbol{\gamma}^*$ is estimated by minimizing

$$\lambda \left(\boldsymbol{\gamma}^{*T}, \boldsymbol{\gamma}^T\right) \boldsymbol{S}^* \begin{pmatrix} \boldsymbol{\gamma}^* \\ \boldsymbol{\gamma} \end{pmatrix} \tag{5.16}$$

with respect to $\boldsymbol{\gamma}^*$, where $\boldsymbol{\gamma}$ is fixed and $\boldsymbol{S}^*$ is defined as $\boldsymbol{S}$ in (4.12) or in (4.19) with additional rows and columns for the additional parameters $\gamma_l^*$, $l = 0, \ldots, q^* - 1$.

The continuous extension of the parameters, which is a consequence of (5.16), is in the case of a difference penalty greatly influenced by the difference order $k$. For the difference order $k$ the continuous extension of the parameters takes the form of a polynomial with degree $k - 1$.

This can be illustrated for the example of a difference order of $k = 2$. In this case the matrix $\boldsymbol{S}^*$ takes the form $\boldsymbol{D}_2^T \boldsymbol{D}_2$, where

$$\boldsymbol{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}$$

and therefore (5.16) is minimal if

$$\Delta^2 \gamma^*_{l+1} = \gamma^*_{l+1} - 2\gamma^*_l + \gamma^*_{l-1} = 0, \quad l = 1, \ldots, q^* - 2.$$

This is the case if a linear relation of the parameters like $\gamma^*_l = s + tl$ is considered, because

$$\Delta^2 \gamma^*_{l+1} = s + t(l+1) - 2s - 2tl + s + t(l-1) = 0.$$

This corresponds to a continuous linear extension of the model parameters or a polynomial of degree $k - 1 = 1$. But if we would consider a difference order of $k = 3$, a continuous quadratic extension of the model parameters would ensue and so on.

Finally, we want to note that there is no overall criterion which difference order should be used. While for example a difference order of two is the default in R for cubic B-splines, for some data a different difference order could yield to better results. In this context we encourage to try different difference orders and to choose the one which represents the mean behaviour of the data best.

## 5.5. Extensions to multiple Cases

Here we want to address the issue how the model changes if more than one smooth function is considered. For a set of observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{x}^1 = (x^1_1, \ldots, x^1_n)^T$ and $\boldsymbol{x}^2 = (x^2_1, \ldots, x^2_n)^T$ a generalized additive model with two smooth functions can be written as

$$g(\boldsymbol{\mu}) = f_1(\boldsymbol{x}^1) + f_2(\boldsymbol{x}^2), \tag{5.17}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are smooth functions in the sense of continuous functions with continuous first and second derivatives, $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{y}]$ and $g(\cdot)$ represents a monotone link function.

From the claim that the smooth functions $f_1(\cdot)$ and $f_2(\cdot)$ can be written as

$$f_k(\boldsymbol{x}^k) = \sum_{j=0}^{q_k - 1} \gamma^k_j b_j(\boldsymbol{x}^k), \qquad k = 1, 2,$$

it follows that they are linear in the parameters and can therefore be represented by

$$f_k(\boldsymbol{x}^k) = \boldsymbol{X}_k \boldsymbol{\gamma}^k, \qquad k = 1, 2,$$

where $\boldsymbol{\gamma}^k = (\gamma^k_0, \ldots, \gamma^k_{q_k - 1})^T$.

As a result, the model formula takes the form

$$\begin{aligned} g(\boldsymbol{\mu}) &= \boldsymbol{X}_1 \boldsymbol{\gamma}^1 + \boldsymbol{X}_2 \boldsymbol{\gamma}^2 = \left(\boldsymbol{X}_1, \boldsymbol{X}_2\right) \begin{pmatrix} \boldsymbol{\gamma}^1 \\ \boldsymbol{\gamma}^2 \end{pmatrix}, \\ &= \boldsymbol{X}\, \boldsymbol{\gamma}, \end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{1\,T}, \boldsymbol{\gamma}^{2\,T})^T$.

While the model formula can, thus, be written similar as before, the question how the penalty changes with additional smooth functions arises. Until now all penalties, derivative or difference, could be written as

$$\lambda\, \boldsymbol{\gamma}^T\, \boldsymbol{S}\, \boldsymbol{\gamma},$$

where $\lambda$ describes the smoothing parameter. Since the penalty assures that the smooth function is not too wiggly we need a penalty for each smooth functions, i.e.

$$\lambda_1\, \boldsymbol{\gamma}^{1\,T}\, \tilde{\boldsymbol{S}}_1\, \boldsymbol{\gamma}^1 + \lambda_2\, \boldsymbol{\gamma}^{2\,T}\, \tilde{\boldsymbol{S}}_2\, \boldsymbol{\gamma}^2,$$

where $\tilde{\boldsymbol{S}}_1$ is a $q_1 \times q_1$ matrix while $\tilde{\boldsymbol{S}}_2$ has dimension $q_2 \times q_2$. By adding zeros to the matrices, we get two $q \times q$ matrices $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, where $q = q_1 + q_2$, and the resulting penalty is

$$\lambda_1\, \boldsymbol{\gamma}^T\, \boldsymbol{S}_1\, \boldsymbol{\gamma} + \lambda_2\, \boldsymbol{\gamma}^T\, \boldsymbol{S}_2\, \boldsymbol{\gamma} = \boldsymbol{\gamma}^T\, (\lambda_1\, \boldsymbol{S}_1 + \lambda_2\, \boldsymbol{S}_2)\, \boldsymbol{\gamma}\,.$$

In the literature, for example in Wood (2006a), the weighted sum of the penalty matrices is often represented by one penalty matrix, meaning

$$\lambda_1\, \boldsymbol{S}_1 + \lambda_2\, \boldsymbol{S}_2 = \boldsymbol{S}^*,$$

and the penalty therefore reduces to

$$\boldsymbol{\gamma}^T\, \boldsymbol{S}^*\, \boldsymbol{\gamma}\,.$$

Therefore, if the smoothing parameter $\lambda$ is included in the penalty matrix $\boldsymbol{S}^*$, we get the same penalty as before. In addition, the estimation of $\lambda_1$ and $\lambda_2$ works as described in 5.2 by minimizing the UBRE or the GCV criterion.

Next, we want to discuss a model with an additional parametric term, i.e

$$g(\boldsymbol{\mu}) = \boldsymbol{x}^0\, \gamma_0 + f_1(\boldsymbol{x}^1) + f_2(\boldsymbol{x}^2),$$

where $\boldsymbol{x}^0 = (x_1^0, \ldots, x_n^0)^T$ represents another explanatory variable. Since we did already establish that $f_1(\boldsymbol{x}^1) + f_2(\boldsymbol{x}^2)$ can be written as $(\boldsymbol{X}_1, \boldsymbol{X}_2) \begin{pmatrix} \boldsymbol{\gamma}^1 \\ \boldsymbol{\gamma}^2 \end{pmatrix}$, it follows that

$$\begin{aligned} g(\boldsymbol{\mu}) &= \boldsymbol{x}^0\, \gamma_0 + (\boldsymbol{X}_1, \boldsymbol{X}_2) \begin{pmatrix} \boldsymbol{\gamma}^1 \\ \boldsymbol{\gamma}^2 \end{pmatrix} \\ &= (\boldsymbol{x}^0, \boldsymbol{X}_1, \boldsymbol{X}_2) \begin{pmatrix} \gamma_0 \\ \boldsymbol{\gamma}^1 \\ \boldsymbol{\gamma}^2 \end{pmatrix} = \boldsymbol{X}\, \boldsymbol{\gamma}\,. \end{aligned}$$

Therefore, by adding the column $\boldsymbol{x}^0$ to $\boldsymbol{X}$ and including $\gamma_0$ in $\boldsymbol{\gamma}$, a parametric term can be added to the model. Since the parametric term is equivalent to a linear fit, we do not need a penalty for this part (no wiggliness). Then the penalty matrix corresponding

to this model is given by $\boldsymbol{S}^* = \lambda_1 \boldsymbol{S}_1 + \lambda_2 \boldsymbol{S}_2$ and by adding one row and one column of zeros to $\boldsymbol{S}^*$ because of the additional parameter in $\boldsymbol{\gamma}$.

Summarizing, for each additional predictor $\boldsymbol{x}_k$ a new column is added to the model matrix $\boldsymbol{X}$, while the penalty matrix $\boldsymbol{S}$ is extended by a row and a column of zeros. On the other hand, an additional smooth function $f_k(\cdot)$ adds the number of the basis functions $q_k$ as columns to the model matrix $\boldsymbol{X}$, while the penalty is extended for $\lambda_k \boldsymbol{S}_k$.

We want to note that here no interaction between different smooth functions is considered. If one wants to allow for interactions, instead of one dimensional basis functions two dimensional basis functions need to be considered and a smooth surface is estimated instead of two smooth functions. More information on this topic can be found for example in Wood (2006a) or in Currie and Durban (2002).

# 6. Practical Example: The Gas Flow Data

In this chapter we want to apply the theory of the last chapters to a practical example. For this case we use the gas flow dataset provided by Open Grid Europe GmbH (OGE), see Friedl et al. (2012), which was already mentioned earlier. While until now we have only considered the variables `temp` and `max.flow` of the dataset `pday`, describing the mean temperature on one day and the daily maximal gas flow, respectively, now the binary indicator `day` distinguishing between working days and weekends and holidays will be used additionally. Thus we have

- `temp` ... the daily mean temperature in degree Celsius,
- `max.flow` ... the daily maximal gas flow in KWh per hour, and
- `day` ... the type of day defined as the binary factor

$$\texttt{day}_i = \begin{cases} 1 & \text{if the } i\text{-th day is a working day} \\ 0 & \text{if the } i\text{-th day is a weekend or holiday} \end{cases}, \; i = 1, \ldots, n.$$

For a better understanding the data is plotted in Figure 6.1. While on the x-axis the temperature is shown, the y-axis represents the maximal gas flow. In addition, the colour of the data points corresponds to the type of day. In other words, the blue colour describes working days, while red stands for weekends and holidays. One can observe that the two groups of data show a similar mean behaviour on working days as on weekends and holidays.

In the following sections models are fitted to the data and confidence and prediction intervals are provided. Finally, the issue of extrapolation is addressed and discussed for the example of the gas flow data.

## 6.1. Working Days versus Weekends and Holidays

In this part we want to estimate a generalized additive model for the gas flow data. While we can observe in Figure 6.1 that the data behaves similar on working days and on weekends and holidays, the question arises which model should be considered. Is it best to disregard the variable `day` and estimate one smooth function or to consider the difference between working days and weekends and holidays by a constant shift?

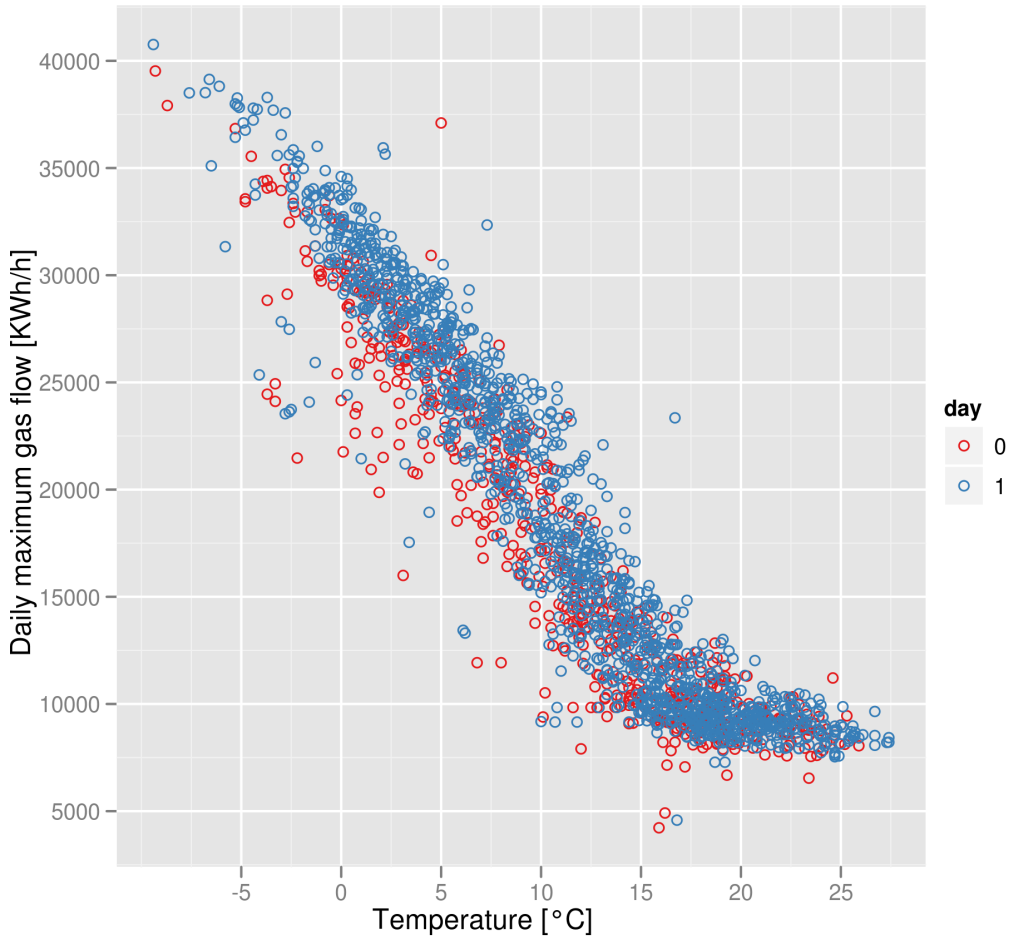


Figure 6.1.: Daily maximum gas flow depending on the respective daily mean temperature and on the type of day (blue - working days, red - weekends and holidays).

On the other hand, we can estimate separate smooth functions for working days and for weekends and holidays. In the following, we will address these different options and discuss them. But before we start by defining the different models which will be discussed in the course of this chapter, we want to call attention to the fact that our responses are maxima. Therefore, a normal distribution might not be appropriate. Nonetheless we at first discuss the respective models under the assumption of a normal distribution and compare the results later on to the fits resulting under a gamma distribution, which is more suited in this case.

In detail the models we will consider for $y_i \sim \mathcal{N}(\mu_i, \phi)$ are:

- `m1`: The first mean model consists of one smooth function of the temperature.

$$\begin{aligned} \mathbb{E}[y_i] &= s(\text{temp}_i) \\ \text{var}(y_i) &= \phi \end{aligned}$$

- `m2`: In addition to the variable in the first model the second includes the factor day.

$$\begin{aligned} \mathbb{E}[y_i] &= \text{day}_i + s(\text{temp}_i) \\ \text{var}(y_i) &= \phi \end{aligned}$$

- `m3`: In the third model the smooth function interacts with the type of day, which results in two smooth functions.

$$\begin{aligned} \mathbb{E}[y_i] &= \text{day}_i * s(\text{temp}_i) \\ \text{var}(y_i) &= \phi \end{aligned}$$

- `m4`: Finally, two separate models each including one smooth function of the temperature are fitted.

$$\begin{aligned} \mathbb{E}[y_i] &= \text{day}_i * s(\text{temp}_i) \\ \text{var}(y_i) &= \phi^{\text{day}_i} \end{aligned}$$

Since `m3` and `m4` both estimate a separate smooth function on working days and on weekends and holidays, the mean models are the same. They only differ in the fact that `m4` allows for two different variances, while in `m3` only one global variance is considered.

Next we start by estimating one smooth function for the whole data and therefore neglecting the variable `day` (`m1`). The resulting fit is shown in Figure 6.2, while in the following R-code the estimation of the model is presented. The R-code necessary to produce the plots which are presented in this chapter can be found in Appendix C.

After estimating the model with the command `gam` from the package `mgcv`, the output resulting from the command `summary` is presented. At first the used member of the exponential family is stated, then the link function and the model formula are described. Thereafter, the estimated coefficients and their p-values first for the parametric terms and then for the smooth terms are specified. Although in `m1` no parametric term is specified, we note that an intercept is automatically added. Here the intercept represents the mean of the response variable $\boldsymbol{y}$, i.e.

$$\hat{\beta}_0 = \bar{y}.$$

This adds numerical stability. In addition, the smooth function now describes the deviation from the mean. Finally, the R-squared, the GCV score, the estimate of the dispersion parameter $\phi$ and the number of observations are stated.

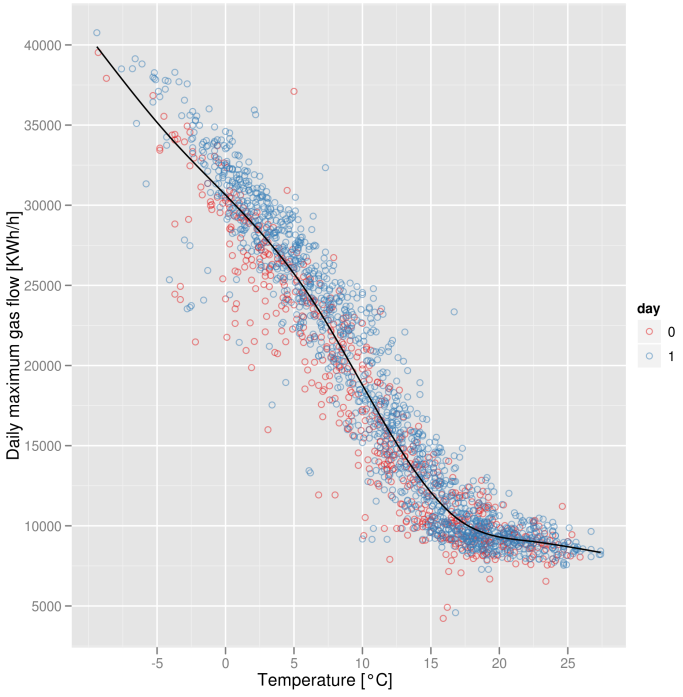


Figure 6.2.: One smooth function resulting from cubic P-splines with difference order two and $q = 10$ (m1).

```
m1<-gam(max.flow~s(temp,bs="ps",m=c(2,2),k=10))
summary(m1)

Family: gaussian
Link function: identity

Formula:
max.flow ~ s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18192.01      52.43     347   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df    F p-value
s(temp) 5.98   6.69 3526  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
R-sq.(adj) =  0.922   Deviance explained = 92.2%
GCV score = 5.5394e+06  Scale est. = 5.5202e+06  n = 2008
```

The p-values in the R-code correspond to the hypothesis tests that the intercept is zero and that the smooth function in the temperature is zero, respectively, i.e.

$$H_0: \quad \beta_0 = 0,$$
$$H_0: \quad s(\text{temp}) = 0.$$

Since we did already establish that the intercept is estimated by the mean of the response variable, testing if the intercept is zero is equivalent to testing if the overall mean is zero. On the other hand, testing if the smooth function is equal to zero is equivalent to testing if all $\boldsymbol{\gamma}$ are zero. If all parameters of the smooth function are zero, the smooth function is a constant function. Since the smooth function describes the deviation from the mean, $\boldsymbol{\gamma} = \mathbf{0}$ means that there is no deviation from the global mean and therefore the smooth function is not needed.

While the fit resulting from this model (see Figure 6.2) seems to describe the data well, the question how the type of day affects it needs to be further discussed. Therefore, in the next model the parametric explanatory factor `day` is included. The following R-code shows the command for model estimation and the output of the resulting model.

```
m2<-gam(max.flow~factor(day)+s(temp,bs="ps",m=c(2,2),k=10))
summary(m2)

Family: gaussian
Link function: identity

Formula:
max.flow ~ factor(day) + s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   17229.8       90.1  191.24   <2e-16 ***
factor(day)1   1401.1      108.8   12.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df    F p-value
s(temp) 6.112  6.792 3758  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.928   Deviance explained = 92.8%
GCV score = 5.1197e+06  Scale est. = 5.099e+06  n = 2008
```

In the R output above the factor day is highly significant, indicating that there is a relevant difference between working days and weekends and holidays. Furthermore, in comparison to the previous model the GCV score and the estimate of the dispersion parameter could be reduced, while also the R-squared is a slightly larger value now.

The two parallel smooth functions resulting from this model are shown in Figure 6.3. As in the previous plots the blue colour represents the working days, while red indicates weekends and holidays. Although the two fits seem to represent the data well enough, on the right margin the red function seems to be too low.

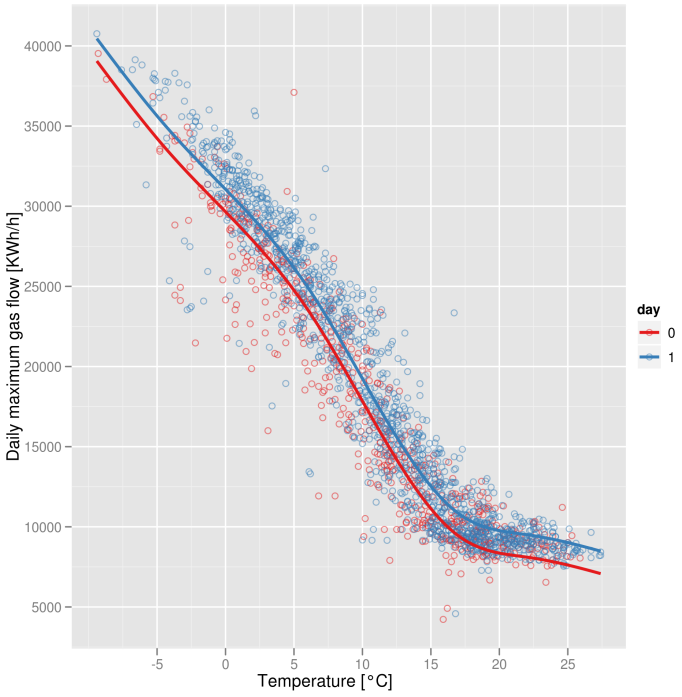


Figure 6.3.: Two smooth functions resulting from cubic P-splines (difference order two, $q = 10$) and from a constant shift parameter (`m2`).

For a better understanding in a next step one smooth function for working days and a separate one for weekends and holidays are fitted to the data (`m3`). This can be achieved by using the property `by`, which allows to multiply an explanatory variable with another in the context of smooth functions, see the R-code below.

```
m3<-gam(max.flow~factor(day)
                +s(temp,bs="ps",m=c(2,2),k=10,by=factor(day)))
summary(m3)

Family: gaussian
Link function: identity
```

```
Formula:
max.flow ~ factor(day) + s(temp, bs = "ps", m = c(2, 2), k = 10,
    by = factor(day))

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17231.23      89.05  193.51   <2e-16 ***
factor(day)1  1399.09     107.37   13.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                        edf Ref.df    F p-value
s(temp):factor(day)0 5.407  6.197 1137  <2e-16 ***
s(temp):factor(day)1 5.821  6.554 2937  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.93   Deviance explained =   93%
GCV score = 4.9929e+06  Scale est. = 4.96e+06  n = 2008
```

In addition to the factor `day` the summary of the model is now based on two different smooth functions one for each type of day. One might also notice that each smooth function is significantly different from zero. Moreover, the GCV score and the estimate for the single constant dispersion parameter could be further reduced while the R-squared is slightly higher. The two smooth functions resulting from this model are shown in Figure 6.4. While they seem parallel for the most part, they approach each other on the right margin until they overlap.

Finally, we are interested in a comparison of the resulting fits of the last two models. Especially the difference between a shift parameter and two separate fits seems intriguing. In Figure 6.5 the respective fits of these models are plotted. While the solid lines represent the model using a shift parameter, the dashed lines correspond to the model estimating two separate smooth functions. One can observe that while the fits show a similar behaviour on the left end of the plot, on the right end the separate fits somehow tend towards each other, while the shift parameter forces the solid lines to keep a constant distance.

As in `m3` we now consider a model with two smooth functions, one for working days and one for weekends and holidays - `m4`. While there is no difference to `m3` regarding the model fits, now the two models have different variances while in `m3` only one parameter $\phi$ is allowed for the entire data. In the R-code below the estimation of `m4` and the output of the command `summary` are presented.

```
#Data
pday0<-subset(pday,day==0) #weekends and holidays
```

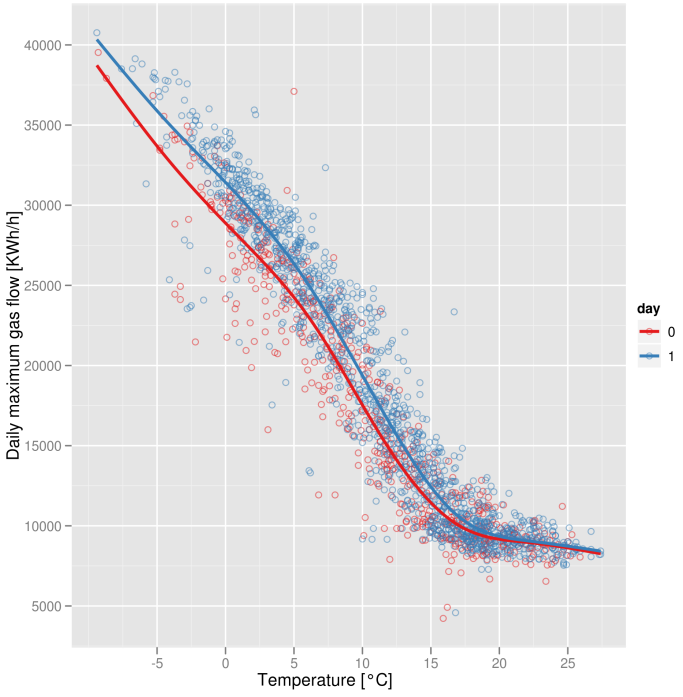


Figure 6.4.: One smooth function for each type of day resulting from cubic P-splines (difference order two, $q = 10$), see m3.

```
pday1<-subset(pday,day==1) #working days

#weekends and holidays
m40<-gam(max.flow~s(temp,bs="ps",m=c(2,2),k=10),data=pday0)
summary(m40)

Family: gaussian
Link function: identity

Formula:
max.flow ~ s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17177.50      92.56   185.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df    F p-value
```

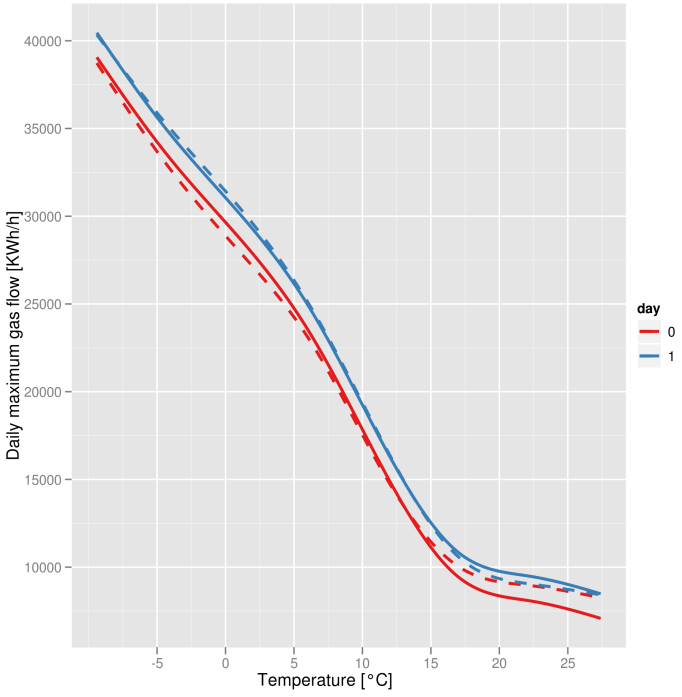


Figure 6.5.: Comparison of the smooth functions resulting from using one shift parameter (solid lines - `m2`) or estimating two separate functions (dashed lines - `m3`) with cubic P-splines (difference order two, $q = 10$).

```
s(temp) 5.387  6.198 1046  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.912   Deviance explained = 91.2%
GCV score = 5.4442e+06  Scale est. = 5.389e+06  n = 629


#working days
m41<-gam(max.flow~s(temp,bs="ps",m=c(2,2),k=10),data=pday1)
summary(m41)

Family: gaussian
Link function: identity

Formula:
max.flow ~ s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18654.75      58.77   317.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df    F p-value
s(temp) 5.856  6.583 3045  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.936   Deviance explained = 93.6%
GCV score = 4.7867e+06  Scale est. = 4.7629e+06  n = 1379
```

Since `m3` and `m4` estimate the same smooth functions and therefore result in the same fits, a comparison of the fits is useless. In addition, the fact that `m4` consists of two models makes model comparison rather difficult (no nested models). However, in Table 6.1 the AIC of the different models is stated for model comparison. The respective AIC values were calculated in R with the command `AIC()`, i.e.

```
AIC(m1)
[1] 36879.46
```

While for the first three models this approach is straight forward, for the fourth model the AIC is calculated by summing the AIC values of the two models `m40` and `m41`. The additivity of the AIC ensures that the resulting AIC values are comparable. From `m1` to `m3` we can observe a gradual improvement of the AIC, while the difference between `m3` and `m4` is very small, which is supported by our previous findings.

| | AIC |
|---|---|
| `m1` | 36 879.46 |
| `m2` | 36 721.22 |
| `m3` | 36 670.81 |
| `m4` | 36 669.09 |

Table 6.1.: Comparison of the AIC for the models `m1` to `m4` under the assumption of a normal distribution.

We want to note that all models in this part were estimated under the assumption of a normal distribution. As to the adequacy of this assumption we want to refer to the next section were the approach based on a normal distribution is compared to using the gamma distribution instead.

## 6.2. Normal versus gamma distribution

While all models in the last part were estimated under the assumption of a normal distribution and an identity link, we now want to consider an alternative model, namely the gamma distribution with a log link. As we did already illustrate at the start of this chapter, a gamma distribution is more appropriate than a normal distribution since we consider maxima. The log link is not the canonical link of the gamma distribution (see Chapter 3). However, it definitely makes more sense in this case, since our responses are positive and a log link ensures this property. Therefore, the option `family` of the command `gam` is now set to `Gamma(link=log)`.

First of all, we take a look at the model using a shift parameter and assuming a gamma distribution. While the following R-code summarizes the output of the command `summary` for this model, the fit is shown in Figure 6.6. In addition to the solid lines representing the current model, the previous model using a shift parameter and assuming a normal distribution is represented by two dotted lines. While the shift parameter in combination with the normal distribution ensures that the fits are parallel, the log link in the current model guarantees positive means and allows them to get closer to each other on the right end (as suggested by `m3` before).

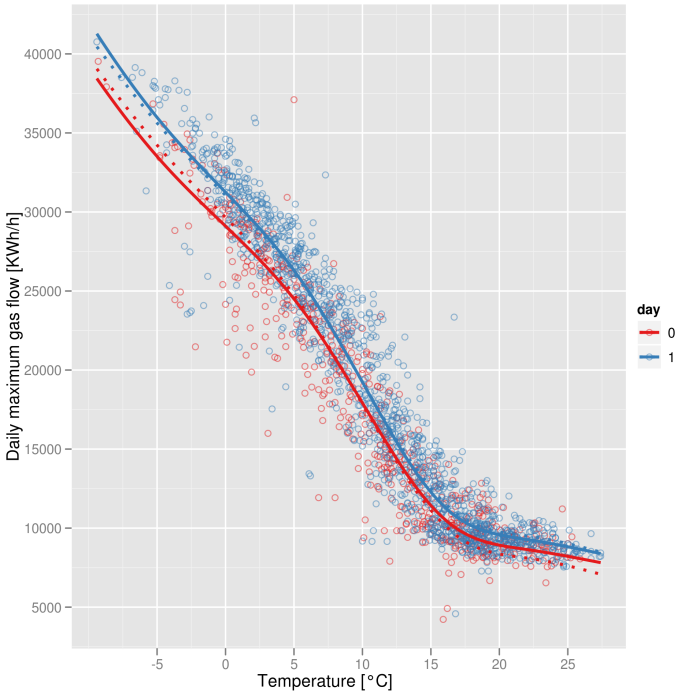


Figure 6.6.: Smooth functions resulting from `m2` using cubic P-splines (difference order two, $q = 10$), where the shape of the lines corresponds to the distribution, i.e. gamma distribution (solid lines) and normal distribution (dotted lines).

```
Family<-"Gamma(link=log)"

m2_g<-gam(max.flow~ factor(day)
               +s(temp,bs="ps",m=c(2,2),k=10),family=Family)
summary(m2_g)

Family: Gamma
Link function: log

Formula:
max.flow ~ factor(day) + s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.657974   0.005246 1841.11   <2e-16 ***
factor(day)1 0.071059   0.006332   11.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df    F p-value
s(temp) 6.499  7.076 3383  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.929   Deviance explained = 92.2%
GCV score = 0.017358  Scale est. = 0.017285  n = 2008
```

In the R output above we notice that for the first time the R-squared and the explained deviance differ. This is a result from the definition of the two terms which is according to the corresponding R help page as follows: The R-squared is defined as the proportion of variance explained, where the variance and the residual variance are calculated using unbiased estimators. To be more precise the adjusted R-squared is defined as

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/(n - \mathrm{tr}(\boldsymbol{A}))}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}, \tag{6.1}$$

where the only difference to the adjusted R-squared in the linear model case is that now the number of parameters $p$ is replaced by the trace of the influence matrix $\boldsymbol{A}$. The computation of the adjusted R-squared in R is described in the R code below, where `object` stands for a model estimated by `gam`.

```
radj<-function(object){
  1-(sum((object$y-object$fitted)^2)/(object$df.res))/
    (sum((object$y-mean(object$y))^2)/object$df.null)
}
```

However, the explained deviance is defined as the proportion of the null deviance explained by the deviance of the respective model. Therefore, the explained deviance can be calculated as

$$D_{explained} = \frac{D(\boldsymbol{y}, \bar{y}) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{D(\boldsymbol{y}, \bar{y})}, \tag{6.2}$$

where $D(\boldsymbol{y}, \bar{y})$ describes the null deviance, which is the deviance of a model including only an intercept. The R-code to calculate the explained deviance is given below. As before the term `object` represents any model estimated by `gam`.

```
dvex<-function(object){
  (object$null.deviance - object$deviance)/object$null.deviance
}
```

In case of a normal distribution the R-squared and the null deviance are nearly the same while for other distributions they do not necessarily need to be. This is the reason why in the last section no difference between the adjusted R-squared and the explained deviance was observed, while they differ now.

Next we consider the model allowing for separate smooth functions for working days and weekends and holidays (`m3`). While the R-code below summarizes the results of this model considering a gamma distribution, the corresponding fit is represented in Figure 6.7. Again the model resulting from the assumption of normal distribution is illustrated by dotted lines, while the solid lines represent the current model assuming a gamma distribution with log link. One might observe that there is practically no difference between the fits of the two models.

```
m3_g<-gam(max.flow~ factor(day)
           +s(temp,bs="ps",m=c(2,2),k=10,by=factor(day)),family=Family)
summary(m3_g)

Family: Gamma
Link function: log

Formula:
max.flow ~ factor(day) + s(temp, bs = "ps", m = c(2, 2), k = 10,
    by = factor(day))

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.658207   0.005234 1845.44   <2e-16 ***
factor(day)1 0.070825   0.006310   11.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

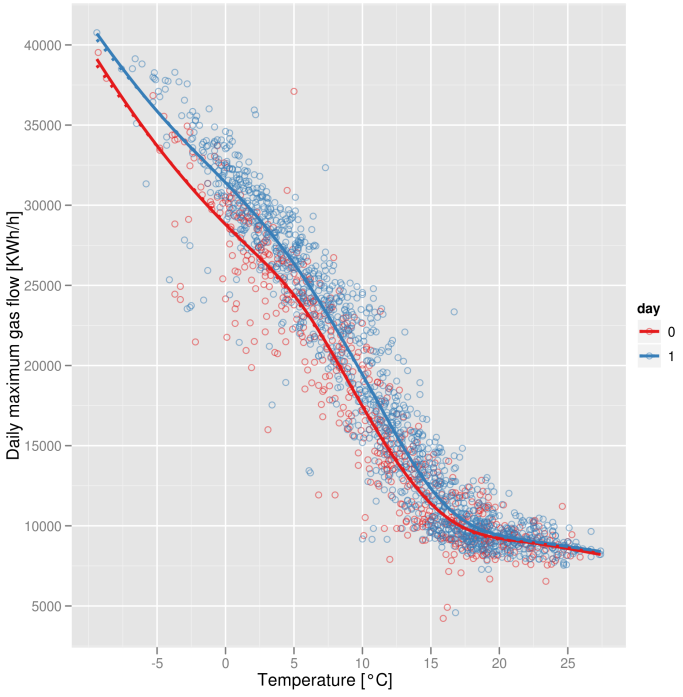


Figure 6.7.: Smooth functions for each type of day resulting from `m3` using cubic P-splines (difference order two, $q = 10$), where the shape of the lines corresponds to the distribution, i.e. gamma distribution (solid lines) and normal distribution (dotted lines).

```
Approximate significance of smooth terms:
                         edf Ref.df    F p-value
s(temp):factor(day)0 5.766  6.499 1074  <2e-16 ***
s(temp):factor(day)1 6.306  6.944 2479  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.93   Deviance explained = 92.3%
GCV score = 0.017244  Scale est. = 0.017123  n = 2008
```

While the difference between the shift parameter model (`m2`) and the model with two separate fits (`m3`) is the same as in the last section, the difference seems to be smaller than before. This can be observed for example in Figure 6.8, where the solid lines correspond to the model using a shift parameter, while the dashed lines represent the model with two smooth functions. In both cases the fits are closer to each other at the right margin than before, but in case of the two separate smooth functions the fits at some point overlap each other, while the shift parameter forces the solid lines to stay apart.

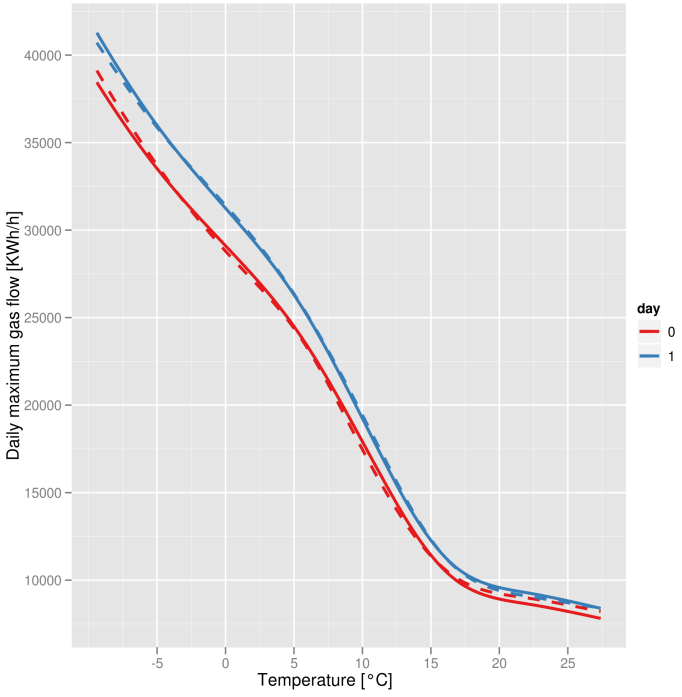


Figure 6.8.: Comparison of the smooth functions resulting from using one shift parameter (solid lines - `m2`) or estimating two separate functions (dashed lines - `m3`) with cubic P-splines (difference order two, $q = 10$) and assuming a gamma distribution with log link.

To evaluate which of the two models should be chosen one can consider the respective analysis of deviance (anova). For GAMs in R the command `anova.gam` can achieve such a task. In the following code the command for an F-test and its results are shown, first for the models assuming a normal distribution then for the models based on a gamma distribution. In both cases the model using two smooth functions is chosen over the model using a shift parameter.

```
anova.gam(m2,m3,test="F")
Analysis of Deviance Table

Model 1: max.flow ~ factor(day) + s(temp, bs = "ps",
                    m = c(2, 2), k = 10)
Model 2: max.flow ~ factor(day) + s(temp, bs = "ps",
                    m = c(2, 2), k = 10, by = factor(day))
  Resid. Df Resid. Dev     Df  Deviance      F    Pr(>F)
1    1999.9 1.0197e+10
2    1994.8 9.8941e+09 5.1158 303341706 11.955 1.242e-11 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


anova.gam(m2_g,m3_g,test="F")
Analysis of Deviance Table

Model 1: max.flow ~ factor(day) + s(temp, bs = "ps",
                    m = c(2, 2), k = 10)
Model 2: max.flow ~ factor(day) + s(temp, bs = "ps",
                    m = c(2, 2), k = 10,
    by = factor(day))
  Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
1    1999.5     34.561
2    1993.9     34.143 5.5729   0.4178 4.3782 0.0003233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Models `m2` and `m3` only differ in the smooth terms, since the parametric term is included in both cases by the factor `day`. While in `m2` only one smooth function of the temperature is considered, in `m3` one smooth function is estimated for each type of day. Therefore, the anova tests if the two smooth functions in `m3` significantly differ from the smooth function in `m2`. Although the p-values result from approximations and should therefore be considered carefully, it seems as if `m3` should be preferred to `m2`.

Since `m4` consists of two models, an analysis of deviance like before is not possible with it. But we did already establish that `m3` and `m4` are basically the same. While they both fit two smooth functions and therefore produce the same fits, `m3` estimates one dispersion parameter and `m4` two. Naturally, `m4` can also be estimated assuming a gamma distribution with log link instead of a normal distribution, see R-code.

```
#working days
m41_g<-gam(max.flow~s(temp,bs="ps",m=c(2,2),k=10),data=pday1,
                  family=Family)
summary(m41_g)

Family: Gamma
Link function: log

Formula:
max.flow ~ s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.729523   0.003394    2867   <2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(temp) 6.35  6.978 2659  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.936   Deviance explained =   93%
GCV score = 0.015971  Scale est. = 0.015886  n = 1379


#weekends and holidays
m40_g<-gam(max.flow~s(temp,bs="ps",m=c(2,2),k=10),data=pday0,
                  family=Family)
summary(m40_g)

Family: Gamma
Link function: log

Formula:
max.flow ~ s(temp, bs = "ps", m = c(2, 2), k = 10)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.656548   0.005612    1721   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(temp) 5.687  6.461 934.3  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.912   Deviance explained = 90.6%
GCV score = 0.020022  Scale est. = 0.019809  n = 629
```

Comparing the AIC of the models in Table 6.2 yields to similar results as in case of a normal distribution earlier. While an improvement is observable from `m1` to `m3`, the difference between `m3` and `m4` is rather small. Furthermore, the best AIC value under the assumption of a normal distribution is worse than the worst AIC value under the assumption of a gamma distribution. Indicating that a gamma distribution should be clearly preferred to a normal distribution.

| | AIC |
|---|---|
| m1_g | 36 634.32 |
| m2_g | 36 514.20 |
| m3_g | 36 500.86 |
| m4_g | 36 490.98 |

Table 6.2.: Comparison of the AIC for the models m1 to m4 under the assumption of a gamma distribution with log link.

Nonetheless, the question remains if m3 or m4 should be preferred. Since we will address confidence and prediction intervals next, we favour m4 over m3. In addition, the calculation of the model matrix $\boldsymbol{X}$ is easier for m4 which will be favourable for extrapolation later.

## 6.2.1. Confidence and Prediction Intervals

Now the confidence and prediction intervals for m4 are generated and compared dependent on the assumed response distribution.

### Confidence Intervals

In order to start we provide the R-code necessary to calculate confidence intervals. Results are shown in Figure 6.9. The confidence intervals are calculated using the command predict.gam and the findings of Chapter 5 regarding the delta method. But before we estimate the model m40, we set up the model m40_s. It is the same model as m40 but the option fit=FALSE allows us to set up the model only but not to fit it. As a consequence, we can derive useful information from m40_s like for example the model matrix $\boldsymbol{X}$. If we want to estimate the model m40_s in a next step, we specify G=m40_s in gam, which performs the estimation and we thereby get m40.

The rest of the R-code is pretty straight forward. Using predict.gam we estimate the standard deviation of the mean $\hat{\boldsymbol{\mu}}$ and summarize it and the fit in P0. Thereafter, we calculate the upper and lower bound of the confidence interval and combine them in the variable rib10, where 0 is due to the used data, namely pday0. In a next step we calculate the variance of the mean with the delta method. On that account we define the first derivative of the inverse link function for our choice of distribution and call it d_linkinv. Then we calculate the variance of the mean Vf0 using the delta method and the upper and lower bound of the confidence interval, which we summarize in rib20.

```
Family<-"Gamma(link=log)" #Family<-"gaussian"
pday0<-subset(pday,day==0) #Weekends and holidays
pday1<-subset(pday,day==1) #working days

alpha<-0.05
```

```
m40_s<-gam(max.flow~s(temp,bs="ps",m=c(2,2),k=10),data=pday0,
                  family=Family,fit=FALSE)
m40<-gam(G=m40_s)

#gam-predict
P0<-predict.gam(m40,type="response",se.fit=TRUE)

up<-P0$fit - qt(1-alpha/2,m40$df.residual)*P0$se.fit
low<-P0$fit + qt(1-alpha/2,m40$df.residual)*P0$se.fit
rib10<-data.frame(pday0,up,low,fit1=P0$fit)

#delta method
if(Family=="gaussian") d_linkinv<-function(eta) {rep(1,length(eta))}
  else d_linkinv<-function(eta) {as.numeric(exp(eta))}

Vf0<-diag(d_linkinv(m40_s$X %*% m40$coef)) %*%
     (m40_s$X %*%  m40$Ve %*% t(m40_s$X)) %*%
     diag(d_linkinv(m40_s$X %*% m40$coef))

up<-m40$fitted - qt(1-alpha/2,m40$df.residual)*sqrt(diag(Vf0))
low<-m40$fitted + qt(1-alpha/2,m40$df.residual)*sqrt(diag(Vf0))
rib20<-data.frame(pday0,up,low,fit1=m40$fitted)
```

Using the R-code above but substituting `pday0` by `pday1` leads to `m41` and its fitted values and confidence intervals respectively. Hence we get `rib11` and `rib21` by substituting `pday0` by `pday1` and `m40` by `m41`. After combining the data in the R-code below the code to generate the plots in Figure 6.9 is given.

```
#combine data
rib1<-rbind(rib10,rib11)
rib2<-rbind(rib20,rib21)

#plot
ggplot(rib1,aes(temp,fit1,colour=as.factor(day),group=factor(day))) +
  geom_line(size=1) +
  geom_ribbon(aes(ymin=low,ymax=up,fill=as.factor(day)),data=rib1,
              alpha=1/3)  +
  geom_ribbon(aes(ymin=low,ymax=up,fill=as.factor(day)),data=rib2,
              alpha=1/3,linetype=2)
```

In Figure 6.9 two plots showing the confidence intervals under the normal and under the gamma distribution, respectively, are included. One might observe that the confidence intervals resulting from the command `predict.gam` (solid lines) and the confidence intervals resulting form the delta method (dashed lines) are very similar but not

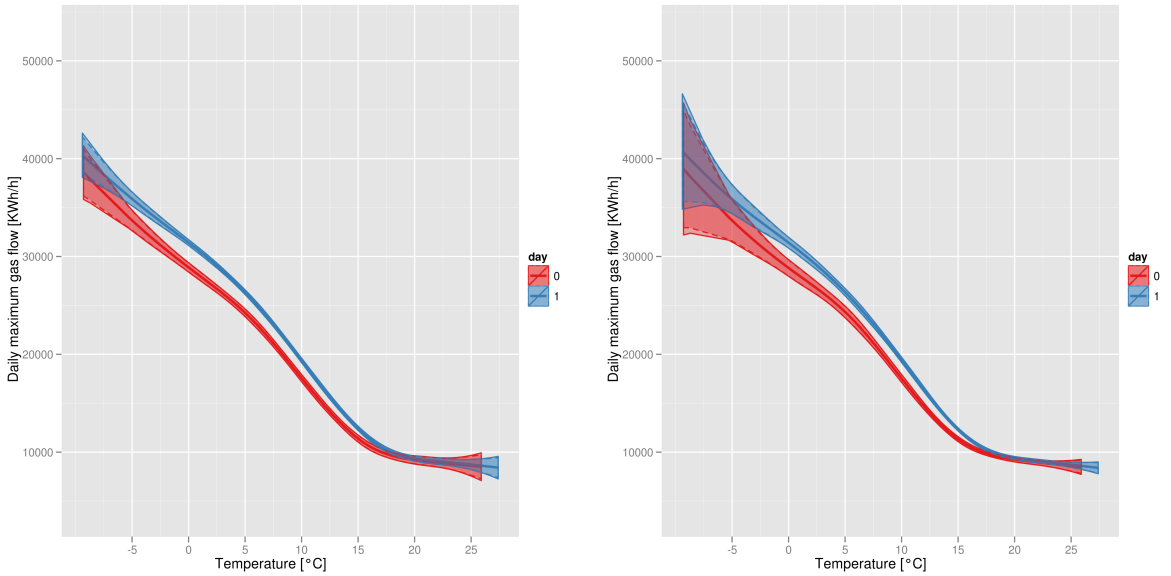


Figure 6.9.: Confidence intervals ($\alpha = 0.05$) using `predict.gam` (solid lines) or using the delta method (dashed lines) for cubic P-splines (difference order two, $q = 10$) and assuming a normal distribution, identity link model (left plot) or a gamma distribution, log link model (right plot).

the same. A noticeable difference is only distinguishable in the upper left corner of the plots, where the confidence intervals grow wide. In addition, we notice the difference between the normal and the gamma distribution. While previously concerning the fit there was no noticeable difference, the confidence intervals show differences especially in the upper left corner. In this case one can observe that the gamma distribution allows for a higher variance as the normal distribution.

### Prediction Intervals

Next we want to take a look at prediction intervals for new observations. As shown in Chapter 5 the variance of the difference between a new observation $y^*$ and its mean is given as

$$\text{var}\left(y^* - \hat{\mu}^*\right) = \phi V(\mu^*) + \text{var}(\hat{\mu}^*),$$

where the variance of $\hat{\mu}^*$ can be obtained from a call of the command `predict.gam` or through the delta method. In the following R-code the variance of the residual is calculated using one of the already mentioned methods for the variance of $\mu^*$ and then adding the variance of $y^*$. To calculate the variance of $y^*$ we need the function $V(\cdot)$, which can be derived from the model for example by `m40$family$variance`. Thereafter, the already calculated variances of the mean are used to derive the upper and lower bound of the prediction intervals.

Finally, the R-code to plot the prediction intervals is added. In the left plot of Figure 6.10 the prediction interval can be observed for working days if normally distributed responses are considered, while in the right plot the prediction interval assuming a gamma distribution is shown. The same plots for weekends and holidays are shown in Figure 6.11.

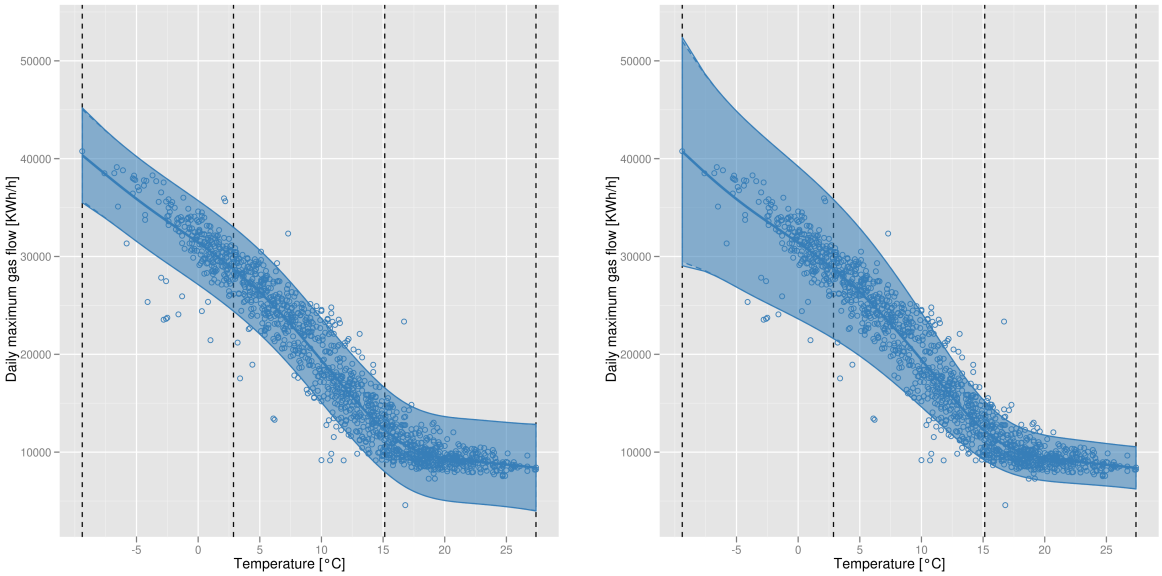


Figure 6.10.: Prediction intervals ($\alpha = 0.05$) for working days using `predict.gam` (blue solid lines) or using the delta method (blue dashed lines) for cubic P-splines (difference order two, $q = 10$) and assuming a normal distribution (left plot) or a gamma distribution (right plot), while the black dashed lines divide the temperature range in three thirds.

```
var_y0<-m40$sig2 * m40$family$variance(rib10$fit)

alpha<-0.05

#gam-predict
rib10$pred_low<-rib10$fit1 - qt(1-alpha/2,m40$df.residual)*
                  sqrt(P0$se.fit^2 + var_y0)
rib10$pred_up<-rib10$fit1 + qt(1-alpha/2,m40$df.residual)*
                  sqrt(P0$se.fit^2 + var_y0)

#delta method
rib20$pred_low<-rib20$fit - qt(1-alpha/2,m40$df.residual)*
                  sqrt(diag(Vf0) + var_y0)
rib20$pred_up<-rib20$fit + qt(1-alpha/2,m40$df.residual)*
                 sqrt(diag(Vf0) + var_y0)
```

```
#plot
col<-"#E41A1C"
p0<-ggplot(pday0,aes(temp,max.flow))
points<-geom_point(colour=col,shape=1)

p0+points +geom_line(aes(y=fit1),data=rib20,size=1,colour=col) +
  geom_ribbon(aes(ymin=pred_low,ymax=pred_up),data=rib10,fill=col,
                alpha=1/3,colour=col)  +
  geom_ribbon(aes(ymin=pred_low,ymax=pred_up),data=rib20,fill=col,
                alpha=1/3,colour=col,linetype=2)
```

| | Normal distribution | Gamma distribution |
|---|---|---|
| Whole data | 0.949 | 0.943 |
| First third | 0.940 | 0.974 |
| Second third | 0.920 | 0.926 |
| Last third | 0.994 | 0.953 |

Table 6.3.: Percentage of the data inside the prediction interval for $\alpha = 0.05$ on working days.

| | Normal distribution | Gamma distribution |
|---|---|---|
| Whole data | 0.948 | 0.957 |
| First third | 0.872 | 0.989 |
| Second third | 0.935 | 0.945 |
| Last third | 0.992 | 0.959 |

Table 6.4.: Percentage of the data inside the prediction interval for $\alpha = 0.05$ on weekends and holidays.

While in the left plot of Figure 6.10 the prediction interval is smallest around 10°C, in the right plot the variance is proportional to the mean and therefore increases with larger mean values and is smallest between 20°C and 25°C, which is due to the choice of a gamma model. In addition, we want to mention that the difference between `predict.gam` (solid lines) and the delta method (dashed lines) is rather small. These observations can also be made in Figure 6.11.

In Table 6.3 the coverage percentage of the prediction interval on working days for the whole data range and for the first third (left end), the second third and the last third (right end) is given. The thirds are construed by dividing the temperature range evenly in three parts and can be observed in Figure 6.10 and in Figure 6.11 (black dashed lines). While the overall coverage probability is good, it looks as if the normal distribution is better in the first third and the gamma distribution is more appropriate in the last third.

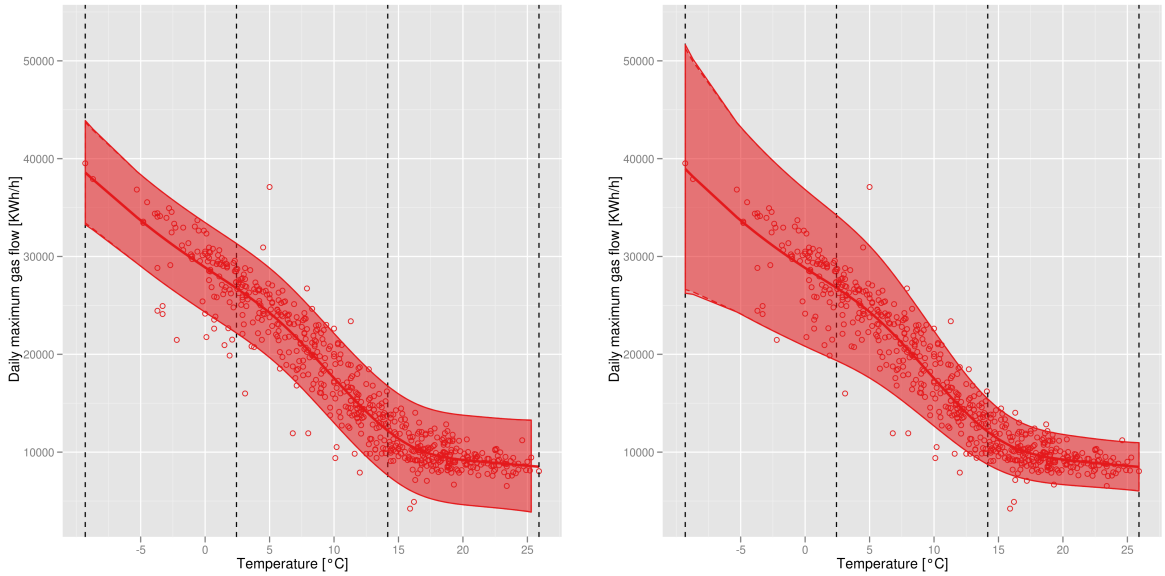


Figure 6.11.: Prediction intervals ($\alpha = 0.05$) for weekends using `predict.gam` (red solid lines) or using the delta method (red dashed lines) for cubic P-splines (difference order two, $q = 10$) and assuming a normal distribution (left plot) or a gamma distribution (right plot), while the black dashed lines divide the temperature range in three thirds.

This is due to the fact that in case of a normal distribution we assume that $\text{var}(y) = \phi$, the variance stays constant, which results in an underestimation of the variance in the first third and an overestimation in the last third. On the other hand, the assumption of a gamma distribution with $\text{var}(y) = \phi\mu^2$ results in an overestimation of the variance in the first third, while the performance in the second and last third is okay.

The coverage percentage of the prediction interval on weekends and holidays is summarized in Table 6.4. It is possible to make the same observation as for working days: We notice a good overall coverage but also a difference between the three thirds. In addition, the gamma distribution performs well in the last third while the normal distribution results in a too narrow prediction interval in the first third and a too wide one in the last third.

The assumption of a gamma model seems reasonable since we are looking at daily maxima and it therefore might seem appropriate to assume that the variance increases with higher observed values rather than staying constant. On the other hand, the gamma distribution clearly (see Figure 6.10 and Figure 6.11 as well as Table 6.3 and 6.4) overestimates the variance in the first third. Furthermore, the data shows no increase of the variance proportional to the mean. To the contrary it makes much more sense that at a certain level saturation is reached, meaning that the use of gas for heating has reached its maximum and no further increase in dependence of the temperature is possible, which results in comparatively little variance for low temperatures.

While the assumption of an increasing variance might not be appropriate, there is still

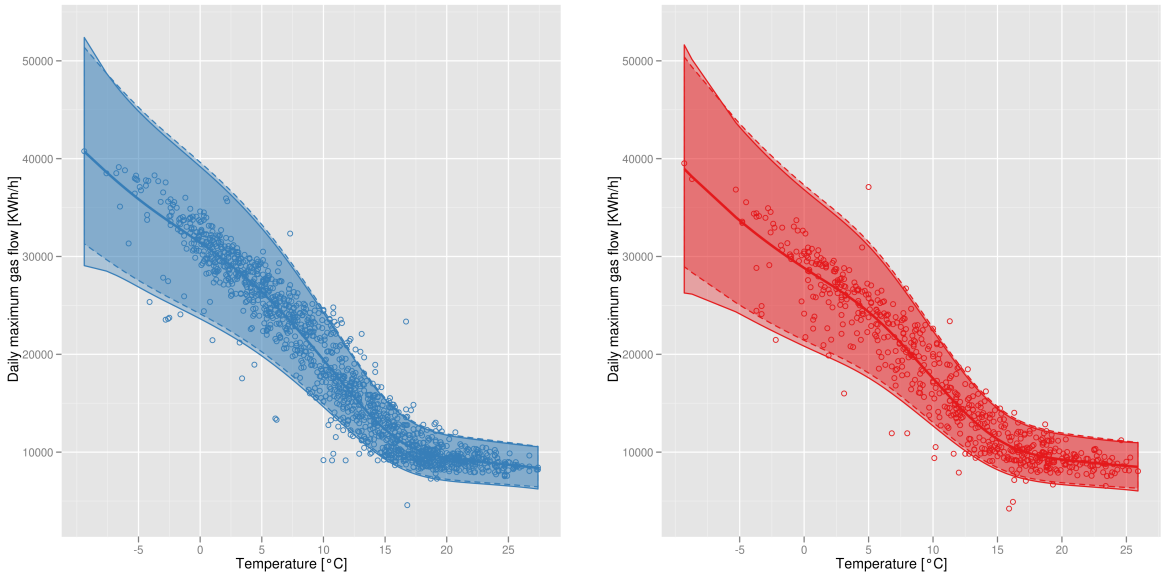


Figure 6.12.: Prediction intervals ($\alpha = 0.05$) for working days (left plot) and weekends and holidays (right plot) using `predict.gam` (solid lines) or using quantiles of the gamma distribution (dashed lines) for cubic P-splines (difference order two, $q = 10$) and assuming a gamma distribution.

the possibility that the data is not symmetrically distributed. Since we did assume a symmetrical distribution around the mean for our previous prediction intervals, we need to address this issue now. Therefore, we calculate prediction intervals as quantiles of a gamma distribution using the estimated mean and variance. Thereby we allow this new prediction intervals to be asymmetric, which would be observable in comparison to the previous symmetric prediction intervals.

In the R-code below we define the function `gamma_q`, which returns the quantile of a gamma distribution using the estimated model and $\alpha$ as input parameters. With this function new prediction intervals can be calculated and compared to the previous ones. In Figure 6.12 this comparison can be observed. A real difference between the prediction intervals can only be noticed in the first third, where the prediction interval using the assumption of a symmetrical distribution around the mean is larger. This is due to the fact that this prediction interval includes the variance resulting from the estimation of the mean, while the prediction interval based on quantiles of the gamma distribution is not factoring that in. Especially in the first third where the variance of the mean is higher due to fewer observations, this leads to a difference between the two intervals. In regard of symmetry no great deviation from it can be observed. Therefore, we conclude that our symmetric prediction intervals are valid.

```
gamma_q<-function(alpha,object){
    mu<-object$fitted
    var<-object$sig2 * object$family$variance(mu)
```

```
    a<-mu^2/var #shape
    s<-var/mu #scale
    qgamma(alpha,shape=a,scale=s) #quantile
}

#weekends and holidays
rib10$pred_lows<-gamma_q(alpha/2,m40)
rib10$pred_ups<-gamma_q(1-alpha/2,m40)

#working days
rib11$pred_lows<-gamma_q(alpha/2,m41)
rib11$pred_ups<-gamma_q(1-alpha/2,m41)

#plot
col<-"#E41A1C"
p0<-ggplot(pday0,aes(temp,max.flow))
points<-geom_point(colour=col,shape=1)

p0+points +geom_line(aes(y=fit1),data=rib10,size=1,colour=col) +
  geom_ribbon(aes(ymin=pred_low,ymax=pred_up),data=rib10,
                  fill=col,alpha=1/3,colour=col)  +
  geom_ribbon(aes(ymin=pred_lows,ymax=pred_ups),data=rib10,
                  fill=col,alpha=1/3,colour=col,linetype=2)
```

Summarizing we note that while the gamma distribution performs well especially in the last third, the prediction intervals under the assumption of a normal distribution seem to represent the data (see Figure 6.3 and Figure 6.4) better. Especially the assumption of a quadratic increase of the variance in respect of the mean is not supported by the data. On the contrary it seems more plausible that for high temperatures and for low temperatures the smallest variance should be observed due to the fact that at a certain state either all radiators are on or off. Therefore, in the following section only a normal distribution will be considered.

## 6.3. Extrapolation

Finally, we want to address the problem of predicting a mean response if the temperature is as low as $-15°C$. To answer this we need to extrapolate the estimated fit, which is accomplished by using the information of Section 5.4. Similar as in Currie et al. (2003) we minimize the difference between the new parameters resulting from the new basis functions and the already estimated model parameters, although in our case `gam` performs the minimization.

In the following R-code we describe one way to extrapolate the fit. Later we will show an easier way to do it with `predict.gam`, but since both ways yield to the same result,

the first way helps to understand the process. The code starts by defining the extended dataset, which includes temperatures as low as $-15°C$. Since we intend to extrapolate from model `m4`, two datasets are defined - one for working days and one for weekends and holidays. Below the R-code to extrapolate the mean behaviour on weekends and holidays is added. But changing `pday0` to `pday1` leads to the extrapolated mean on working days.

Next we define the number of inner knots `K`, the degree of the B-splines `m` (`m=3` for cubic B-splines) and the used member of the exponential family. Thereafter, the B-splines are provided. In a next step we compute the penalty matrix `S` and get everything we need to estimate the model. To do that we use the option `paraPen`, which allows us to penalize a parametric variable. Since the columns of `Bs_f` denote the B-splines and each column is equivalent to a parametric variable, the option ensures that the B-splines are penalized. Although for the first B-spline no responses are available, the model estimates the corresponding parameter. Since this is a consequence of the minimization of the penalty, we will see later that in case of a difference order of two this corresponds to a continuous linear extension of the parameters like showed in Section 5.4.

Finally, the extrapolation and the prediction interval are calculated using the already estimated parameters and the delta method for the variance of the mean. The derivation of the prediction intervals is the same as in Section 6.2.1.

```
#new data
pday_ex<-data.frame(flow.date=rep(NA,2*6),temp=rep(-15:-10,2),
                      max.flow=rep(NA,2*6),day=rep(0:1,each=6))
pday_n<-rbind(pday_ex,pday) #new dataset

pday0<-subset(pday_n,day==0) #weekends and holidays
pday1<-subset(pday_n,day==1) #working days
attach(pday0)          #alternativ: attach(pday1)

K<-6;m<-3
Family<-"gaussian"

Bs_f<-splineDesign(knots((-m):(K+m+1),K,temp),temp,ord=m+1) #B-splines
q<-dim(Bs_f)[2] #number of B-splines

d<-2 #difference order
P<-diff(diag(q),differences=d) #d order differences
S<-t(P)%*%P # Penalty matrix

#estimate model
m4t<-gam(max.flow ~ Bs_f,paraPen=list(Bs_f=list(S)),family=Family)

#extrapolation
X_f<-cbind(rep(1,length(Bs_f[,1])),Bs_f)
```

```
B$fit<-X_f %*% m4t$coefficients #for normal distribution
B$se.fit<-sqrt(rowSums((X_f %*% m4t$Vp) * X_f))*
                  abs(m4t$family$mu.eta(X_f %*% m4t$coefficients)) # sd

#prediction interval
var_y<-m4t$sig2 * m4t$family$variance(B$fit)
B$up<-B$fit + qt(1-alpha/2,m4t$df.residual)*sqrt(B$se.fit^2 + var_y)
B$down<-B$fit - qt(1-alpha/2,m4t$df.residual)*sqrt(B$se.fit^2 + var_y)
```

In Section 5.4 we did already show that in case of P-splines with difference order two, the penalty is minimized by a linear extension of the parameters. In the R-code below we analyse this context at the current example. One can observe that the parameter of the first B-spline can be deduced by a linear extension of the second and third parameter.

```
#coef[1] if d=2 on working days
par<-m4t$coefficients[-1]
par[1]
   Bs_f1
  28684.92
par[2] + (par[2]-par[3])
  28684.84
```

An alternative way to extrapolate the mean is illustrated in the next R-code. Here we start by defining the same knots as we used previously to construct the B-splines and then estimate the model with this information. With `predict.gam` we will be able to extrapolate the mean and estimate the standard deviation. Finally, the prediction interval can be deduced in the same way as before. At this point we want to mention that the prediction interval calculated earlier and the prediction interval calculated with `predict.gam` below are identical.

Next we call attention to the knots. For the extrapolation to work we need to specify the knots in `gam` which are used to construct all the B-splines, including the B-splines used for extrapolation. This point is crucial because otherwise the extrapolation as introduced here will not work.

```
#knots
kn<-data.frame(temp=knots((-m):(K+m+1),K,temp)) #define knots

#estimate model
m4<-gam(max.flow~ s(temp,bs="ps",k=q),data=pday0,knots=kn,family=Family)

#extrapolation
A<-predict.gam(m4,newdata=data.frame(temp=pday0$temp),se=TRUE,
                  type="response")
```

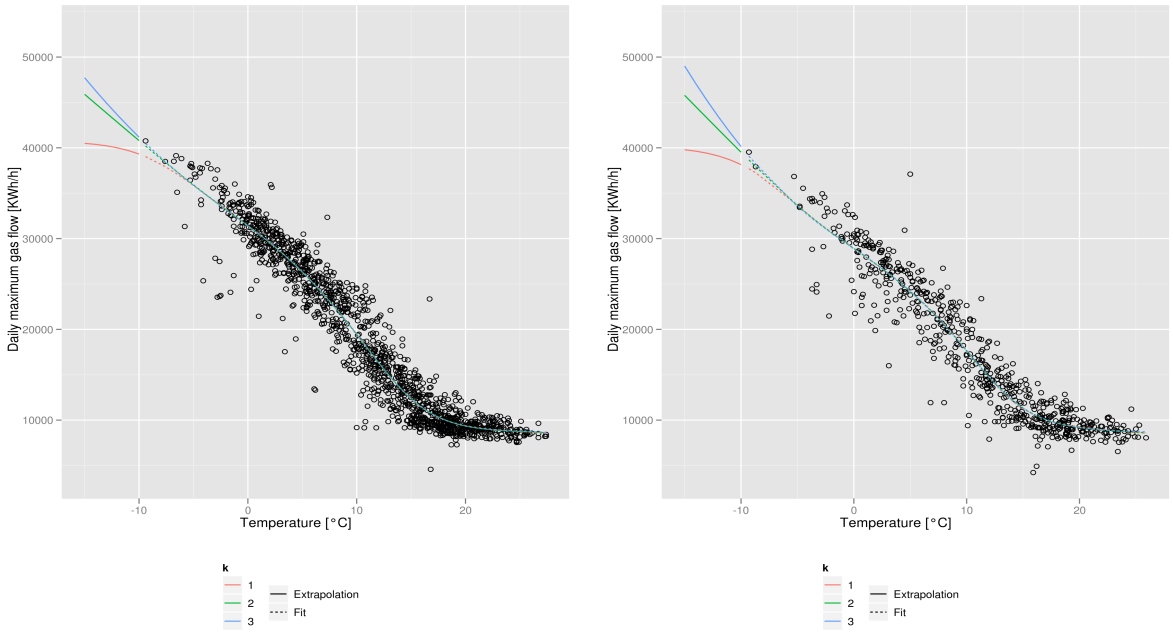


Figure 6.13.: Extrapolation of the mean for working days (left plot) and weekends and holidays (right plot) using cubic P-splines with difference order $k = 1, 2, 3$, $q = 10$ and assuming a normal distribution.

```
#prediction interval
var_y<-m4$sig2 * m4$family$variance(A$fit)
A$up<-A$fit + qt(1-alpha/2,m4$df.residual)*sqrt(A$se.fit^2 + var_y)
A$down<-A$fit - qt(1-alpha/2,m4$df.residual)*sqrt(A$se.fit^2 + var_y)
```

Finally, the results of this extrapolation can be observed in Figure 6.13, where the extrapolation is represented by solid lines while the dashed lines correspond to the smoothly estimated mean. The colour of the lines describes which difference order was used for model estimation and to get the extrapolation. Therefore, slight changes in the mean are also possible if we consider different difference orders. In our case they only seem to occur in the upper left corner of the plots. In addition, there seems to be practically no difference between the extrapolations on working days (left plot) and weekends and holidays (right plot).

Depending on the difference order the extrapolation is a constant, linear or quadratic continuous extension of the mean. Therefore, the mean behaviour of the extrapolated mean is strongly influenced by the choice of the difference order. In our case the most likely continuation of the mean could be between the difference orders one and two.

The prediction intervals resulting from the previous findings can be observed in the left plot of Figure 6.14 for working days and in the right plot for weekends and holidays. In both figures the lighter ribbons on the left side stand for the extrapolation while the ribbon on the right side of the plots describes the prediction interval of the fit.

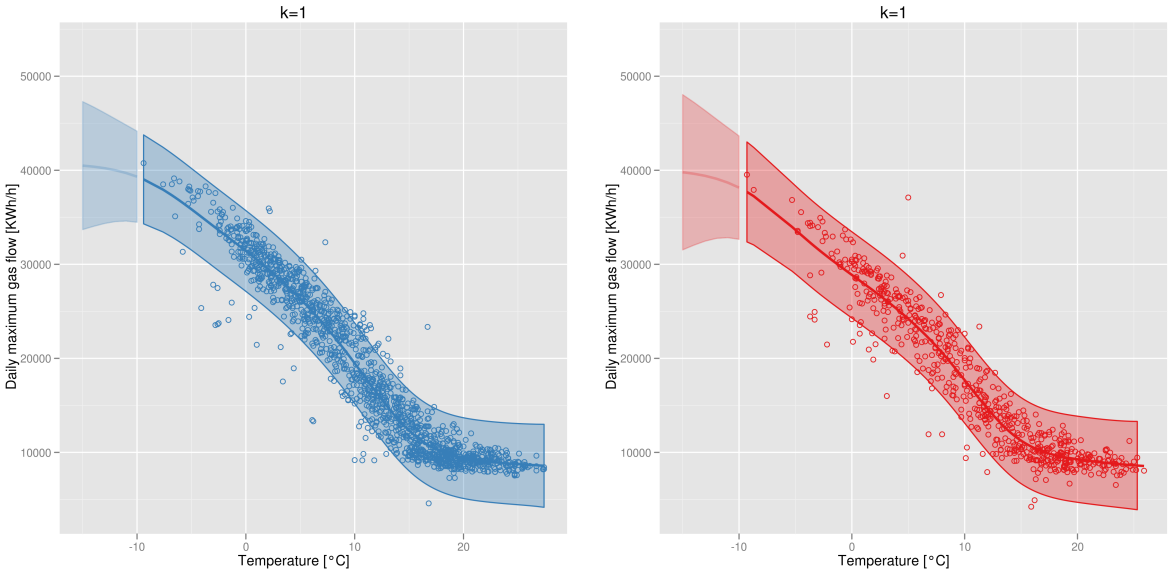


Figure 6.14.: Prediction interval ($\alpha = 0.05$) for the fit and the extrapolation using `predict.gam` for cubic P-splines (difference order one, $q = 10$) on working days (left plot) and on weekends and holidays (right plot), assuming a normal distribution.

The difference order one is selected for this final figure because it seems in our case the best choice. A linear or quadratic increase of the gas flow is very unlikely, since at some point no further increase is technically possible (network limitations). Therefore an asymptote as used in Friedl et al. (2012) makes sense. As a result, the best GAM in our opinion is the one with difference order one shown in Figure 6.14

# 7. Summary

The challenge of this work was to estimate the mean of the daily gas flows and to extrapolate it for colder temperatures. Since the daily maximal gas flow depends in a nonlinear way on the daily mean temperature, first additive models and thereafter generalized additive models were applied. In Chapter 6 we did conclude that a normal distribution is better suited to the data than a gamma distribution with a log link, which would have overestimated the variance for higher response values.

In addition, the influence of the factor `day`, which distinguishes between working days and weekends and holidays, was analysed in Chapter 6. While a constant shift parameter performs quite well, a difference in the mean behaviour on working days and on weekends and holidays turned out to be significant. As a result, the mean was estimated separately on working days and on weekends and holidays.

In a final step the mean was extrapolated to temperatures below $-10°C$. Since no observations are available there, the fit is greatly dependent on the selected difference order (in case of P-splines with a difference order penalty). To mark these uncertainties confidence and prediction intervals where calculated and shown in plots.

# A. First derivative of B-splines

Here we want to proof that the first derivative of a B-spline defined by

$$B_j^m(x) = \frac{x - \kappa_j}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) + \frac{\kappa_{j+m+1} - x}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x),$$

where $j = -m, \ldots, K$, and

$$B_j^0(x) = \begin{cases} 1 & \kappa_j \leq x \leq \kappa_{j+1} \\ 0 & \text{otherwise} \end{cases} \qquad j = 0, \ldots, K,$$

can be written as

$$\frac{\partial B_j^m(x)}{\partial x} = m \left( \frac{1}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) - \frac{1}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x) \right),$$

where $m$ describes the degree of the spline and $K$ the number of inner knots and $x \in [\kappa_j, \kappa_{j+1}]$.

The proof is done by induction on $m$. For $m = 1$ the first derivative for $x \in [\kappa_j, \kappa_{j+1}]$ is

$$\begin{aligned} \frac{\partial B_j^1(x)}{\partial x} =& \frac{1}{\kappa_{j+1} - \kappa_j} B_j^0(x) + \frac{x - \kappa_j}{\kappa_{j+1} - \kappa_j} \frac{\partial B_j^0(x)}{\partial x} \\ & - \frac{1}{\kappa_{j+2} - \kappa_{j+1}} B_{j+1}^0(x) + \frac{\kappa_{j+2} - x}{\kappa_{j+2} - \kappa_{j+1}} \frac{\partial B_{j+1}^0(x)}{\partial x}. \end{aligned}$$

From the definition of $B_j^0(x)$ above it follows that the first derivative of $B_j^0(x)$ is zero for $x \in [\kappa_j, \kappa_{j+1}]$ and

$$\frac{\partial B_j^1(x)}{\partial x} = \frac{1}{\kappa_{j+1} - \kappa_j} B_j^0(x) - \frac{1}{\kappa_{j+2} - \kappa_{j+1}} B_{j+1}^0(x),$$

which is our induction basis.

In the induction step from $m-1$ to $m$ we get that

$$\begin{aligned}
\frac{\partial B_j^m(x)}{\partial x} &= \frac{1}{\kappa_{j+m}-\kappa_j} B_j^{m-1}(x) + \frac{x-\kappa_j}{\kappa_{j+m}-\kappa_j}\frac{\partial B_j^{m-1}(x)}{\partial x} \\
&\quad - \frac{1}{\kappa_{j+m+1}-\kappa_{j+1}} B_{j+1}^{m-1}(x) + \frac{\kappa_{j+m+1}-x}{\kappa_{j+m+1}-\kappa_{j+1}}\frac{\partial B_{j+1}^{m-1}(x)}{\partial x} \\
&= \frac{1}{\kappa_{j+m}-\kappa_j} B_j^{m-1}(x) \\
&\quad + \frac{x-\kappa_j}{\kappa_{j+m}-\kappa_j}(m-1)\left(\frac{1}{\kappa_{j+m+1}-\kappa_j} B_j^{m-2}(x) - \frac{1}{\kappa_{j+m}-\kappa_{j+1}} B_{j+1}^{m-2}(x)\right) \\
&\quad - \frac{1}{\kappa_{j+m+1}-\kappa_{j+1}} B_{j+1}^{m-1}(x) \\
&\quad + \frac{\kappa_{j+m+1}-x}{\kappa_{j+m+1}-\kappa_{j+1}}(m-1)\left(\frac{1}{\kappa_{j+m}-\kappa_{j+1}} B_{j+1}^{m-2}(x) - \frac{1}{\kappa_{j+m+1}-\kappa_{j+2}} B_{j+2}^{m-2}(x)\right)
\end{aligned}$$

In the equation above the term $B_{j+1}^{m-2}(x)$ appears two times. Before we continue we need to rearrange the parameters of this term, therefore

$$\begin{aligned}
&- \frac{x-\kappa_j}{\kappa_{j+m}-\kappa_j}\frac{1}{\kappa_{j+m}-\kappa_{j+1}} + \frac{\kappa_{j+m+1}-x}{\kappa_{j+m+1}-\kappa_{j+1}}\frac{1}{\kappa_{j+m}-\kappa_{j+1}} \\
&= \frac{-(x-\kappa_j)(\kappa_{j+m+1}-\kappa_{j+1}) + (\kappa_{j+m+1}-x)(\kappa_{j+m}-\kappa_j)}{(\kappa_{j+m}-\kappa_j)(\kappa_{j+m}-\kappa_{j+1})(\kappa_{j+m+1}-\kappa_{j+1})} \\
&= \frac{-x\kappa_{j+m+1} + x\kappa_{j+1} + \kappa_j\kappa_{j+m+1} - \kappa_j\kappa_{j+1} + \kappa_{j+m+1}\kappa_{j+m} - \kappa_{j+m+1}\kappa_j - x\kappa_{j+m} + x\kappa_j}{(\kappa_{j+m}-\kappa_j)(\kappa_{j+m}-\kappa_{j+1})(\kappa_{j+m+1}-\kappa_{j+1})} \\
&= \frac{-x\kappa_{j+m+1} + x\kappa_{j+1} + \kappa_{j+m+1}\kappa_{j+m} - \kappa_{j+m}\kappa_{j+1} + \kappa_{j+m}\kappa_{j+1} - \kappa_j\kappa_{j+1} - x\kappa_{j+m} + x\kappa_j}{(\kappa_{j+m}-\kappa_j)(\kappa_{j+m}-\kappa_{j+1})(\kappa_{j+m+1}-\kappa_{j+1})} \\
&= \frac{(\kappa_{j+m+1}-\kappa_{j+1})(\kappa_{j+m}-x) - (\kappa_{j+m}-\kappa_j)(x-\kappa_{j+1})}{(\kappa_{j+m}-\kappa_j)(\kappa_{j+m}-\kappa_{j+1})(\kappa_{j+m+1}-\kappa_{j+1})} \\
&= \frac{1}{\kappa_{j+m}-\kappa_j}\frac{\kappa_{j+m}-x}{\kappa_{j+m}-\kappa_{j+1}} - \frac{1}{\kappa_{j+m+1}-\kappa_{j+1}}\frac{x-\kappa_{j+1}}{\kappa_{j+m}-\kappa_{j+1}}.
\end{aligned}$$

As a result, we get

$$\begin{aligned}
\frac{\partial B_j^m(x)}{\partial x} &= \frac{1}{\kappa_{j+m}-\kappa_j} B_j^{m-1}(x) \\
&\quad + \frac{1}{\kappa_{j+m}-\kappa_j}(m-1)\left(\frac{x-\kappa_j}{\kappa_{j+m+1}-\kappa_j} B_j^{m-2}(x) + \frac{\kappa_{j+m}-x}{\kappa_{j+m}-\kappa_{j+1}} B_{j+1}^{m-2}(x)\right) \\
&\quad - \frac{1}{\kappa_{j+m+1}-\kappa_{j+1}} B_{j+1}^{m-1}(x) \\
&\quad - \frac{1}{\kappa_{j+m+1}-\kappa_{j+1}}(m-1)\left(\frac{x-\kappa_{j+1}}{\kappa_{j+m}-\kappa_{j+1}} B_{j+1}^{m-2}(x) + \frac{\kappa_{j+m+1}-x}{\kappa_{j+m+1}-\kappa_{j+2}} B_{j+2}^{m-2}(x)\right).
\end{aligned}$$

Therefore, the derivative of a B-spline for $x \in [\kappa_j, \kappa_{j+1}]$ is given by

$$\begin{aligned}\frac{\partial B_j^m(x)}{\partial x} &= \frac{1}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) + (m-1)\frac{1}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) \\ &\quad - \frac{1}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x) - (m-1)\frac{1}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x) \\ &= m\left(\frac{1}{\kappa_{j+m} - \kappa_j} B_j^{m-1}(x) - \frac{1}{\kappa_{j+m+1} - \kappa_{j+1}} B_{j+1}^{m-1}(x)\right).\end{aligned}$$

# B. Proof of Theorem 2

*Proof.* This proof is taken from Wood (2006a) and is based on the Central Limit Theorem of Lindberg and the Chebyshev inequality.

First of all, we define

$$a_i = \sum_j c_j x_{ij} \, \boldsymbol{w}_i,$$

where $x_{ij}$ stands for the respective element of $\boldsymbol{X}$ and $\boldsymbol{w}_i$ describes the $i$-th row of $\boldsymbol{W}$, therefore

$$\mathbf{c}^T \mathbf{v} = \sum_{i=1}^{n} a_i z_i,$$

where $\mathbf{v} = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{z}$.

Next, we set

$$s_n^2 = \sum_{i=1}^{n} a_i^2 \frac{\phi}{w_{ii}},$$

where $w_{ii}$ describe the $i$-th diagonal element of $\boldsymbol{W}$. Next we define

$$\mathcal{U}_i = \begin{cases} a_i z_i - a_i \mu_i & \text{if} \quad |a_i z_i - a_i \mu_i| \leq \varepsilon s_n \\ 0 & \text{if} \quad |a_i z_i - a_i \mu_i| > \varepsilon s_n \end{cases}, \; i = 1, \ldots, n \quad \forall \varepsilon > 0.$$

The Central Limit Theorem of Lindberg states that if

$$\lim_{n \to \infty} \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left[\mathcal{U}_i^2\right] = 1, \tag{B.1}$$

and $s_n \to \infty$ as $n \to \infty$, then

$$\frac{1}{s_n} \sum_{i=1}^{n} a_i \left(z_i - \mu_i\right)$$

converges to the standard normal distribution $\mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$.

Provided that $s_n \to \infty$ as $n \to \infty$ and boundedness of the $a_i$'s, then (B.1) is satisfied, if

$$\mathbb{P}\left[\mathcal{U}_i = 0\right] \xrightarrow{n \to \infty} 0.$$

Since $\mathbf{c}$ is a vector of length $q$ and does not change with increasing $n$, the boundedness of the $a_i$'s is given by the boundedness of $x_{ij}\,\boldsymbol{w}_i$. Therefore, the condition above is met, if

$$\forall \varepsilon \quad \lim_{n\to\infty} \mathbb{P}\left[|a_i z_i - a_i \mu_i| > \varepsilon \sum_{i=1}^{n} \frac{a_i^2 \phi}{w_{ii}}\right] = 0.$$

Taking the Chebyshev inequality into account, it follows that

$$\mathbb{P}\left[|a_i z_i - a_i \mu_i| > \varepsilon \sum_{i=1}^{n} \frac{a_i^2 \phi}{w_{ii}}\right] < \frac{a_i^2 \phi / w_{ii}}{\left(\sum_{i=1}^{n} a_i^2 \phi / w_{ii}\right)^2 \varepsilon^2}.$$

Therefore, for condition (B.1) to hold, we need

$$\frac{a_i^2 / w_{ii}}{\left(\sum_{i=1}^{n} a_i^2 / w_{ii}\right)^2} \xrightarrow{n\to\infty} 0.$$

Since $\mathbf{c}$ is the result of weighed sums over columns of $\boldsymbol{X}$, to fulfil the requirement above it suffices if

$$\frac{\left(x_{ij}\,\boldsymbol{w}_i^{1/2}\right)^2}{\left(\sum_{i=1}^{n}\left(x_{ij}\,\boldsymbol{w}_i^{1/2}\right)^2\right)^2} \to 0 \quad \forall i, j.$$

Wood (2006a) offers an interpretation for this last condition: If no element of $\boldsymbol{X}$ dominates the fit as $n \to \infty$, then $\mathbf{c}^T\mathbf{v}$ converges to a multivariate normal distribution. □

# C. R-code of plots

In the following the R-code for the figures in Chapter 6 is added.

```
farbe<-scale_colour_brewer("day",palette="Set1")

p<-ggplot(pday,aes(temp,max.flow))  +xlab("Temperature [C]") +
       ylab("Daily maximum gas flow [KWh/h]") +farbe
points<-geom_point(aes(colour=factor(day)),shape=1,alpha=I(1/2))

temp_10<-seq(min(temp),max(temp),by=1/10)
fit_p1<-predict.gam(m1,newdata=list(temp=temp_10),
                    type="response")

#one smooth function
p +points +geom_line(aes(x=temp_10,y=fit_p1))


#shift parameter
newdata1<-data.frame(temp=rep(temp_10,2),day=c(rep(0,length(temp_10)),
             rep(1,length(temp_10))))
fit_p2<-predict.gam(m2,newdata=newdata1,type="response")

p +points +geom_line(aes(x=temp,y=fit_p2,colour=factor(day)),
               data=newdata1,size=1)


#two separate fits
fit_p3<-predict.gam(m3,newdata=newdata1,type="response")

p +points +geom_line(aes(x=temp,y=fit_p3,colour=factor(day)),
              data=newdata1,size=1)


#shift parameter norm +gamma
fit_shift2<-geom_line(aes(x=temp,y=fit_p2,colour=factor(day)),
                 data=newdata1,size=1,linetype=3)

p +points +geom_line(aes(x=temp,y=fit_p2_gamma,colour=factor(day)),
```

```
               data=newdata1,size=1) +fit_shift2


#separate fits norm +gamma
fit_sepfits<-geom_line(aes(x=temp,y=fit_p3,colour=factor(day)),
                 data=newdata1,size=1,linetype=3)

p +points +geom_line(aes(x=temp,y=fit_p3_gamma,colour=factor(day)),
               data=newdata1,size=1) +fit_sepfits


#shift vs separate fits gamma
fit_shift_gamma<-geom_line(aes(x=temp,y=fit_p2_gamma,
               colour=factor(day)),data=newdata1,size=1,linetype=1)
fit_sepfits_gamma<-geom_line(aes(x=temp,y=fit_p3_gamma,
                 colour=factor(day)),data=newdata1,size=1,linetype=2)

p + fit_shift_gamma +fit_sepfits_gamma


#extrapolation
pl<-qplot(temp,max.flow,data=B_d,shape=1)
pl +geom_line(aes(y=fit,linetype=typ,color=factor(d)))


#extrapolation + prediction interval
A<-data.frame(A)
ggplot(A,aes(temp,fit,colour=as.factor(day))) +geom_line(size=1) +
  geom_point(aes(y=max.flow),shape=1) +
  geom_ribbon(aes(ymin=down,ymax=up,fill=as.factor(day)),alpha=1/3)
```

# References


Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove: Duxbury Press.

Currie, I. D., & Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling*, *2*, 333-349.

Currie, I. D., Durban, M., & Eilers, P. H. C. (2003). Using P-splines to extrapolate two-dimensional Poisson data. In *Proceedings of the 18th International Workshop on Statistical Modelling* (p. 97-102). Leuven, Belgium.

Eilers, P. H. C., & Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89-121.

Eilers, P. H. C., & Marx, B. (2010). Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 637-653.

Fahrmeir, L., & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, *13*, 342-368.

Friedl, H. (2011). *Generalisierte lineare Modelle.* (Lecture Notes, Institute of Statistics, University of Technology Graz)

Friedl, H., Mirkov, R., & Steinkamp, A. (2012). Modeling and forecasting gas flow on exits of gas transmission networks. *International Statistical Review*, *80*, 24-39.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, *39*, 53-74.

Marx, B., & Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, *28*, 193-209.

Mc Cullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org/`

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer. Available from `http://had.co.nz/ggplot2/book`

Wood, S. N. (2006a). *Generalized Additive Models.* Boca Raton: Chapman and Hall/CRC.

Wood, S. N. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics*, *48*, 445-464.