Zsombor Döme, BSc

Generalisierte Additive Modelle zur Analyse der Staatsschuldenkrisen in Schwellenländern

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Finanz- und Versicherungsmathematik

eingereicht an der

Technischen Universität Graz

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

Institut für Statistik

Graz, Mai 2016

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

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 indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

ZUSAMMENFASSUNG

Diese Arbeit beschäftigt sich mit nichtparametrischen Regressionsmodellen, wobei der Fokus auf der Klasse der Generalisierten Additiven Modelle liegt, eine Erweiterung der Generalisierten Linearen Modelle. Der Vorteil dieser Modellklasse ist, dass die erklärenden Variablen nicht mehr rein parametrisch in das Modell eingehen, sondern flexibel als glatte Funktionen in den linearen Prädiktor integriert werden. Nach den theoretischen Überlegungen wird unter Anwendung dieser die Ausfallwahrscheinlichkeit von Schwellenländern modelliert und abschließend eine einjährige Prognose erstellt.

ABSTRACT

This thesis investigates nonparametric regression models, whereby the focus is placed on the class of generalized additive models, which are an extension of generalized linear models. The big advantage of these models is the fact, that the predictor variables are no longer included in a parametric way, by contrast the linear predictor is stated as the sum of smooth functions of the predictor variables. After discussing the theoretical background we estimate the probability of sovereign debt crises for emerging market economies. Afterwards a one-year prediction is made.

Inhaltsverzeichnis

1 Motivation

Staatsschuldenkrisen sind ein sich wiederholendes Phänomen der Geschichte, die häufig zu gesellschaftlichen und politischen Problemen, wie z.B. einer gestiegenen Arbeitslosigkeit, Bankenkollapsen oder sozialen Unruhen, führen. Gut in Erinnerung ist noch die Krise, welche im Jahr 2007 in den Vereinigten Staaten von Amerika als Finanzkrise ihren Ursprung nahm und anschließend in eine weltweite Wirtschaftskrise überging. Um die entstandene Notlage zu bekämpfen, wurden von großen Wirtschaftsnationen und internationalen Organisationen (z.B. IWF) zahlreiche Maßnahmen gesetzt: Die Herabsetzung wichtiger Zinssätze, die Durchführung von Notverstaatlichungen, der Erwerb von Wertpapieren oder die Einrichtung des Sonderfonds Finanzmarktstabilisierung (SoFFin) in Deutschland sind nur einige Beispiele. Durch die getätigten Zwangsmaßnahmen konnte sich zwar die wirtschaftliche Situation vielerorts bald erholen, jedoch verursachten die Konjunkturprogramme eine hohe Verschuldung vieler Länder. Davon tangiert waren insbesondere Irland, Spanien und Griechenland. Letzteres Land etwa, war Anfang 2010 nicht mehr fähig sich am Kapitalmarkt neues Geld zu beschaffen, um die laufenden Kosten zu tragen. Nur durch den sogenannten „Euro-Rettungsschirm“ konnte in diesen Ländern die drohende Zahlungsunfähigkeit abgewendet werden. Die Konsequenzen dieser Staatsschuldenkrisen sind noch bis heute zu spüren.

Um in der Zukunft drohende Staatsschuldenkrisen frühzeitig zu erkennen, wäre es von großem Interesse, ökonomische und politische Bedingungen ausfindig zu machen, welche für deren Auftreten ausschlaggebend sind. Mit genau dieser Problematik befassen sich auch Manasse und Roubini (2005) bzw. Manasse und Roubini (2009). Sie beschränken sich in Ihrer Analyse auf Schwellenländer und versuchen unter Anwendung der „Classification and Regression Tree“ (CART) Methodik die typischen Indikatoren für eine Staatsinsolvenz festzustellen. Die Idee ist nun, anhand der Ergebnisse dieser Arbeiten ein völlig neues Konzept zu entwickeln, welches uns erlaubt, die Ausfallwahrscheinlichkeiten von Schwellenländern zu modellieren bzw. zu prognostizieren. Dabei machen wir uns die Klasse der Generalisierten Additiven Modelle zunutze. Eingeführt von Hastie und Tibshirani (1986) erlaubt sie im Gegensatz zu der Klasse der Generalisierten Linearen Modellen den linearen Prädiktor als Summe von glatten Funktionen darzustellen. Durch diese Erweiterung kann der Einfluss der erklärenden Variable auf die Responsevariable, welche häufig eine nichtlineare Form aufweist, viel exakter wiedergegeben werden. Ein weiterer Punkt, der dafür spricht Generalisierte Additive Modelle für die Modellierung der Eintrittswahrscheinlichkeit von Staatsschuldenkrisen zu gebrauchen, ist die Tatsache, dass die Responsevariable nicht ausschließlich normalverteilt sein muss, sondern eine Verteilung aus der einparametrigen linearen Exponentialfamilie (z.B. Gamma-, Poisson- oder Binomialverteilung) besitzen kann. Weitere Informationen bzw. genauere Details können beispielsweise Wood (2006), Hastie und Tibshirani (1986, 1990), Marx und Eilers (1998) oder Ruppert, Wand und Carroll (2003) entnommen werden.

Wie bereits erwähnt, liegt das Hauptaugenmerk dieser Arbeit darauf, mittels Generalisierten Additiven Modellen die Ausfallwahrscheinlichkeit von Schwellenländern zu modellieren bzw. vorherzusagen. Doch bevor mit der Modellschätzung begonnen werden kann,

ist es von größter Notwendigkeit theoretische Überlegungen anzustellen. Dabei beginnen wir im folgenden Kapitel 2 mit parametrischen Regressionsmodellen, welche als Fundament für die weiteren Modelle betrachtet werden können. Nach der Einführung der Klassischen Linearen Regression fokussieren wir uns auf deren Verallgemeinerung, die Generalisierten Linearen Modelle. An dieser Stelle beschäftigen wir uns unter anderem mit der Herleitung eines Schätzers für den Parametervektor und der Konstruktion eines Maßes für die Anpassungsgüte des Modells. Anschließend widmen wir uns einem Spezialfall der Generalisierten Linearen Modelle, den Logistischen Regressionsmodellen, da diese besonders für unser praktisches Beispiel von großer Wichtigkeit sind. In Kapitel 3 gehen wir zu nichtparametrischen Regressionsmodellen über, wobei zuerst Additive Modelle und im Anschluss Generalisierte Additive Modelle behandelt werden. Einige Fragen, die dabei beantwortet werden, sind z.B. Wie soll die unbekannte glatte Funktion geschätzt werden? Oder, wie glatt soll diese sein? Nachdem die theoretischen Aspekte diskutiert wurden, befassen wir uns in Kapitel 4 mit dem praktischen Teil dieser Arbeit. Hierbei wird mithilfe der vorgestellten Modelle die Eintrittswahrscheinlichkeit von Staatsschuldenkrisen in Schwellenländern kalkuliert und im Anschluss eine einjährige Prognose erstellt. Kapitel 5 fasst die Ergebnisse des praktischen Beispiels zusammen.

Ein sehr wichtiger Aspekt dieser Arbeit ist die Reproduzierbarkeit der mittels Computer durchgeführten Berechnungen, sodass an vielen Stellen der entsprechende Code angegeben wird. Für jegliche computergestützte Kalkulation wird ausschließlich die Software `R` (R Core Team, 2013) verwendet.

2 Parametrische Regression

2.1 Klassische Lineare Regression

Die Klassische Lineare Regression ist ein statistisches Analyseverfahren, welches versucht, den Zusammenhang zwischen einer interessierenden Variable (auch abhängige Variable oder Responsevariable) y und einer oder mehreren erklärenden Variablen (auch unabhängige Variable oder Prädiktorvariable) $x_1, \ldots, x_{p-1}$ zu modellieren. Gesucht ist also diejenige lineare Funktion, welche die Beziehung zwischen der Responsevariable und der Prädiktorvariable optimal beschreibt.

Angenommen es ist eine Stichprobe vom Unfang n gegeben und die Daten liegen in der Form $(x_{i1}, x_{i2}, \ldots, x_{ip-1}, y_i)$, $i = 1, 2, \ldots, n$, vor, dann lässt sich das lineare Modell als

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i = \mathbf{x}_i^t \boldsymbol{\beta} + \epsilon_i$$

schreiben, wobei $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip-1})^t$ der Prädiktorvektor, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^t$ der Parameterverktor und ϵ_i der sogenannte nicht beobachtbare statistische Fehler mit $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ und $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$, ist. Bei der Klassischen Linearen Regression wird zusätzlich angenommen, dass der statistische Fehler normalverteilt ist, d. h. $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Dies impliziert, dass $y_i \overset{ind}{\sim} N(\mathbf{x}_i^t \boldsymbol{\beta}, \sigma^2)$.

Völlig äquivalent dazu kann das Modell in Vektorschreibweise dargestellt werden als

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

mit dem Responsevektor $\mathbf{y} = (y_1, \ldots, y_n)^t$, dem Fehlervektor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^t$ und der $n \times p$ Designmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np-1} \end{pmatrix}.$$

Das Ziel ist nun die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 für die gegebenen Daten zu schätzen. Eine Möglichkeit für die Schätzung des Parametervektors $\boldsymbol{\beta}$ ist mittels der Methode der Kleinsten Quadrate. Dabei wird jene Gerade der Form $\widehat{\mathbb{E}(y_i)} = \hat{\mu}(\mathbf{x}_i) = \mathbf{x}_i^t \hat{\boldsymbol{\beta}}$ (fitted value) gesucht, sodass die $\hat{\mu}_i$ bestmöglich den wahren Beobachtungen y_i entsprechen, d. h die Summe der quadrierten Residuen $r_i = y_i - \hat{\mu}_i$ soll minimiert werden. Der Kleinste Quadrate Schätzer (LSE) $\hat{\boldsymbol{\beta}}$ minimiert also

$$\mathrm{SSE}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}.$$

Als Lösung erhalten wir, falls $\mathbf{X}^t\mathbf{X}$ regulär ist,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}.$$

Um die Varianz σ^2 zu schätzen, bedienen wir uns der Maximum-Likelihood Methode. Da $\mathbf{y}$ der Normalverteilung genügt, gilt für die Log-Likelihoodfunktion

$$\log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\boldsymbol{\beta}))^2.$$

Nun betrachten wir die beiden ersten Ableitungen der Log-Likelihoodfunktion

$$\frac{\partial}{\partial\beta_j}\log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}(y_i - \mu_i), \qquad j = 0, \ldots, p-1$$
$$\frac{\partial}{\partial\sigma^2}\log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i - \mu_i)^2.$$

Bei gleichzeitigem Nullsetzen der beiden Gleichungen erhalten wir die Normalgleichungen und als Ergebnis den Maximum-Likelihood Schätzer (MLE) $\hat{\boldsymbol{\beta}}$, welcher äquivalent zum LSE $\hat{\boldsymbol{\beta}}$ ist, wie auch den gesuchten MLE für die Varianz

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{1}{n}\mathrm{SSE}(\hat{\boldsymbol{\beta}}).$$

Jedoch ist dieser Schätzer nicht erwartungstreu, daher versuchen wir ihn zu verbessern. Mithilfe von $\mathrm{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$ (Beweis siehe Pokropp, 1994) folgt $\mathbb{E}(\mathrm{SSE}(\hat{\boldsymbol{\beta}})) = \sigma^2(n-p)$ und somit ergibt sich als erwartungstreuer Schätzer für die Varianz σ^2

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

Nun illustrieren wir die oben beschriebene Klassische Lineare Regression anhand eines sehr einfachen Beispiels. Die dafür verwendeten Daten stammen aus Venables und Ripley (2002).

Beispiel 2.1. Bei 35 „Scottish hill races“ sind die benötigten Rekordzeiten und Distanzen erhoben worden, deren Werte in Tabelle 2.1 abgebildet werden. Interessiert sind wir dabei am Zusammenhang zwischen der erwarteten Zeit und der Distanz, siehe Abbildung 2.1 (links). Für das Modell ergibt sich nach der Methode der Kleinsten Quadrate folgende geschätzte lineare Funktion:

$$\widehat{\mathbb{E}(\mathrm{Zeit})} = \hat{\mu}(\mathrm{Distanz}) = -4.841 + 8.330 \cdot \mathrm{Distanz},$$

also $\hat{\beta}_0 = -4.841$ und $\hat{\beta}_1 = 8.330$. Diese Gerade wird in Abbildung 2.1 (rechts) veranschaulicht. Wir können den Steigungsparameter $\hat{\beta}_1 = 8.330$ wie folgt interpretieren: Erhöht sich die Distanz um einen Kilometer, so nimmt die durchschnittliche Zeit um 8.330 Minuten zu.

Veranstaltung	Zeit	Distanz
Greenmantle	16.083	2.5
Carnethy	48.350	6.0
Craig Dunain	33.650	6.0
Ben Rha	45.600	7.5
Ben Lomond	62.267	8.0
Goatfell	73.217	8.0
Bens of Jura	204.617	16.0
Cairnpapple	36.367	6.0
Scolty	29.750	5.0
Traprain	39.750	6.0
Lairig Ghru	192.667	28.0
Dollar	43.050	5.0
Lomonds	65.000	9.5
Cairn Table	44.133	6.0
Eildon Two	26.933	4.5
Cairngorm	72.250	10.0
Seven Hills	98.417	14.0
Knock Hill	78.650	3.0
Black Hill	17.417	4.5
Creag Beag	32.567	5.5
Kildcon Hill	15.950	3.0
Meall Ant-Suidhe	27.900	3.5
Half Ben Nevis	47.633	6.0
Cow Hill	17.933	2.0
N Berwick Law	18.683	3.0
Creag Dubh	26.217	4.0
Burnswark	34.433	6.0
Largo Law	28.567	5.0
Criffel	50.500	6.5
Acmony	20.950	5.0
Ben Nevis	85.583	10.0
Knockfarrel	32.383	6.0
Two Breweries	170.250	18.0
Cockleroi	28.100	4.5
Moffat Chase	159.833	20.0

Tabelle 2.1: Scottish hill races

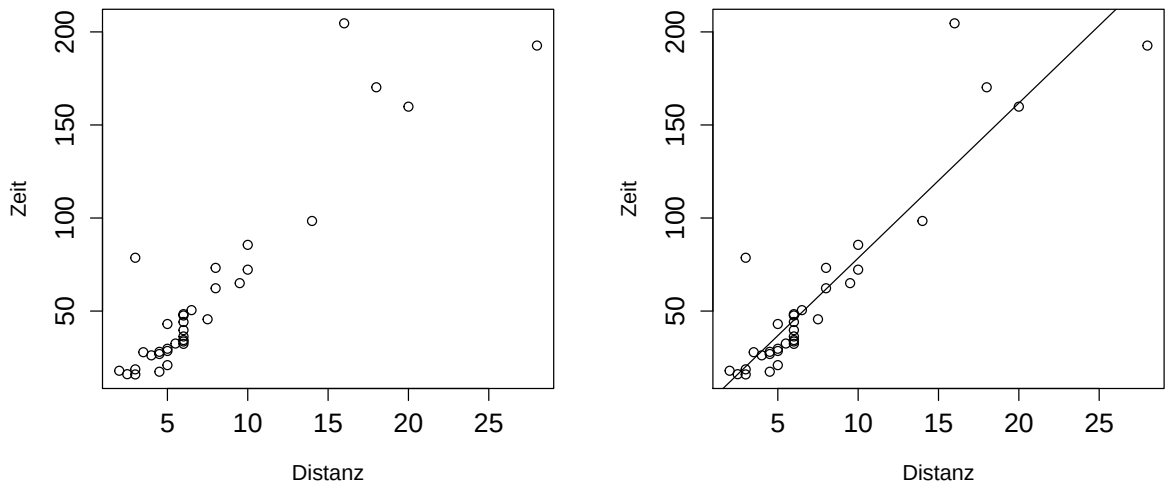

Abbildung 2.1: Streudiagramm zwischen Zeit und Distanz bei „Scottish hill races“ (links). Rechts ist zusätzlich die geschätzte Regressionsgerade eingezeichnet.

2.2 Generalisierte Lineare Modelle

In den folgenden Abschnitten werden einige grundlegende Ideen und Begriffe aus der Theorie der Generalisierten Linearen Modelle bzw. der Logistischen Regression vorgestellt und im Anschluss ein Beispiel zur Illustration dargelegt. Dabei stützen wir uns hauptsächlich auf die Werke McCullagh und Nelder (1989) und Dobson (2001), weitere gute Quellen sind beispielsweise Hardin und Hilbe (2012) und Wood (2006).

Die Klasse der Generalisierten Linearen Modelle (GLM), welche von Nelder und Wedderburn (1972) eingeführt wurde, verallgemeinert die Theorie der Klassischen Linearen Regressionsmodelle. Unterliegt bei der Klassischen Linearen Regression die Responsevariable der Normalverteilung, so stammt diese bei einem GLM aus einer Verteilung der einparametrigen linearen Exponentialfamilie. Diese Erweiterung ist in sehr vielen Bereichen der Wissenschaft nützlich. Werden beispielsweise binäre Responsevariablen (z.B. infiziert oder nicht infiziert, gestorben oder überlebt, etc.) untersucht, kann nicht von einer Normalverteilung ausgegangen werden. Oder wir sind an relativen Anteilen oder Anzahlen interessiert. Weiters kann es vorkommen, dass der Zusammenhang zwischen Erwartungswert und Prädiktorvariable nicht ausschließlich linear ist. Hierfür wird bei einem GLM die sogenannte Linkfunktion $g(\mu_i)$ eingeführt, welche mit dem linearen Prädiktor $\mathbf{x}_i^t\boldsymbol{\beta}$ übereinstimmt.

Mit den folgenden drei Komponente wird die Klasse der Generalisierten Linearen Modelle gebildet (siehe McCullagh und Nelder, 1989):

1. stochastische Komponente: Die Responsevariable unterliegt der einparametrigen linearen Exponentialfamilie, also $y_i \overset{ind}{\sim}$ Exponentialfamilie(θ_i), $i = 1, \ldots, n$.

2. systematische Komponente: $\eta_i = \mathbf{x}_i^t\boldsymbol{\beta}$
   $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^t$ wird als der Vektor mit den linearen Prädiktoren bezeichnet.

3. Linkfunktion: $g(\mu_i) = \eta_i$
   Die Linkfunktion $g(\cdot)$, welche eine bekannte, monotone und zweimal stetig differenzierbare Funktion ist, verbindet den linearen Prädiktor mit dem Erwartungswert.

Im Folgenden widmen wir uns der erweiterten Verteilungsannahme der Responsevariable.

2.2.1 Lineare Exponentialfamilie

Definition 2.1. Eine Zufallsvariable y hat eine Verteilung aus der einparametrigen linearen Exponentialfamilie in kanonischer Form mit kanonischem Parameter θ, falls sich ihre Dichte- oder Wahrscheinlichkeitsfunktion in der folgenden Form darstellen lässt:

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

mit speziellen bekannten Funktionen $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$, wobei $a(\cdot) > 0$ gilt und ϕ bekannt ist.

Nun versuchen wir den Erwartungswert und die Varianz von y für die lineare Exponentialfamilie zu bestimmen. Diese können von den bekannten Eigenschaften (McCullagh und Nelder, 1989)

$$\mathbb{E}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = 0 \tag{2.1}$$

und

$$\mathrm{Var}\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right) = \mathbb{E}\left(\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right)^2\right) = \mathbb{E}\left(-\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2}\right) \tag{2.2}$$

hergeleitet werden. Wird in den beiden Gleichungen die lineare Exponentialfamilie als Verteilung angenommen, so folgt für (2.1)

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

und für (2.2)

$$\mathbb{E}\left(\left(\frac{\partial \log f(y|\theta)}{\partial \theta}\right)^2\right) + \mathbb{E}\left(\frac{\partial^2 \log f(y|\theta)}{\partial \theta^2}\right) = \frac{1}{a^2(\phi)}\,\mathrm{Var}(y) - \frac{1}{a(\phi)}b''(\theta) = 0.$$

Also erhalten wir

$$\begin{aligned} \mathbb{E}(y) &= \mu = b'(\theta), \\ \mathrm{Var}(y) &= a(\phi)b''(\theta). \end{aligned} \tag{2.3}$$

Wir nennen ϕ den Dispersionsparameter sowie $b''(\theta)$ die Varianzfunktion und da diese eine Funktion vom Erwartungswert μ ist, schreiben wir dafür $V(\mu)$. Somit folgt für die Varianz

$$\mathrm{Var}(y) = a(\phi)V(\mu).$$

Für die Dispersionsfunktion $a(\phi)$ beschränken wir uns im Folgenden ausschließlich auf den Fall $a(\phi) = a \cdot \phi$, mit bekanntem Gewicht a.

Als nächstes betrachten wir die Mitglieder der linearen Exponentialfamilie. Zu den wichtigsten gehören unter anderem die Normalverteilung, Poissonverteilung, Binomialverteilung, Gammaverteilung und die Inverse Gaussverteilung. All diese können in kanonischer Form dargestellt werden - siehe Tabelle 2.2.

Verteilung	θ	ϕ	a	$b(\theta)$	$c(y,\phi)$	$V(\mu)$
$y \sim \mathrm{Normal}(\mu,\sigma^2)$	μ	σ^2	1	$\theta^2/2$	$-\frac{1}{2}(\frac{y^2}{\phi} + \log(2\pi\phi))$	1
$y \sim \mathrm{Poisson}(\mu)$	$\log(\mu)$	1	1	$\exp(\theta)$	$-\log y!$	μ
$y \sim \mathrm{Gamma}(\mu,\nu)$	$-1/\mu$	$1/\nu$	1	$-\log(-\theta)$	$\frac{1}{\phi}\log\frac{1}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma(\frac{1}{\phi})$	μ^2
$y \sim \mathrm{InvGauss}(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	σ^2	1	$-(-2\theta)^{1/2}$	$-\frac{1}{2}(\frac{1}{\phi y} + \log(2\pi\phi y^3))$	μ^3
$my \sim \mathrm{Binomial}(m,\pi)$	$\log(\frac{\pi}{1-\pi})$	1	$\frac{1}{m}$	$\log(1+\exp(\theta))$	$\log\binom{m}{my}$	$\mu(1-\mu)$

Tabelle 2.2: Charakteristiken einiger Mitglieder der linearen Exponentialfamilie.

2.2.2 Linkfunktion

Wie bereits erwähnt, verknüpft die Linkfunktion $g(\cdot)$ den linearen Prädiktor η mit dem Erwartungswert μ. Bei der Klassischen Linearen Regression entspricht $g(\cdot)$ der identischen Abbildung. Werden allerdings z.B. Anzahlen behandelt, für welche die Poissonverteilung angenommen wird, so ist der identische Link nicht geeignet, da η negativ sein kann während $\mu > 0$ gelten muss. Für solche Modelle wird in der Regel der Loglink verwendet, $\eta = \log(\mu)$.

Genügen die Responsevariablen einer standardisierten Binomialverteilung, dann gilt $0 < \mu < 1$ und der Link muss die Eigenschaft erfüllen, dass er das Intervall $(0, 1)$ auf die reelle Zahlengerade abbildet. In dieser Modellklasse werden üblicherweise folgende Linkfunktionen verwendet:

- Logitlink:
$$g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right),$$
- Probitlink:
$$g(\mu) = \text{probit}(\mu) = \Phi^{-1}(\mu),$$
wobei $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalverteilung ist.

Generell kann gesagt werden, dass es bei vielen Modellen naheliegend ist, für die Linkfunktion den speziellen Link $g(\mu) = \eta = \theta$ zu wählen. Diese sogenannten kanonischen Links vereinfachen die Modellschätzung und haben vorteilhafte statistische Eigenschaften.

2.2.3 Parameterschätzung

Der folgende Abschnitt diskutiert die Herleitung eines Schätzers für den Parametervektor $\boldsymbol{\beta}$ und orientiert sich an Dobson (2001).

Seien $y_1, \ldots, y_n$ unabhängige Responses aus derselben Exponentialfamilie mit Parameter (θ_i, ϕ), welche die Anforderungen an ein GLM erfüllen. Wir versuchen nun $\boldsymbol{\beta}$ mittels der Maximum-Likelihood Methode zu schätzen. Für die Log-Likelihoodfunktion der Stichprobe ergibt sich

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

Unter Verwendung der Kettenregel der Differentialrechnung und der Annahme $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ gilt für die Scorefunktion

$$\frac{\partial l(\boldsymbol{\theta}|\mathbf{y})}{\partial \beta_j} = U_j = \sum_{i=1}^{n}\left(\frac{\partial l(\theta_i|y_i)}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j}\right), \qquad j = 0, \ldots, p-1.$$

Mit

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

und

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

folgt

$$U_j = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{a_i \phi} \cdot \frac{1}{V(\mu_i)} \cdot x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right) = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\mathrm{Var}(y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right). \tag{2.4}$$

Die Varianz-Kovarianz-Matrix der U_js besitzt die Terme

$$\mathfrak{I}_{jk} = \mathbb{E}(U_j U_k),$$

welche die Informationsmatrix $\boldsymbol{\mathfrak{I}}$ formen. Mit (2.4) ergibt sich

$$\begin{aligned} \mathfrak{I}_{jk} &= \mathbb{E}\left(\sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{\mathrm{Var}(y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^{n} \left[\frac{(y_l - \mu_l)}{\mathrm{Var}(y_l)} x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right) \\ &= \sum_{i=1}^{n} \frac{\mathbb{E}((y_i - \mu_i)^2) x_{ij} x_{ik}}{(\mathrm{Var}(y_i))^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathrm{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{a_i \phi V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \end{aligned} \tag{2.5}$$

Der zweite Schritt erfolgt wegen $\mathbb{E}((y_i - \mu_i)(y_l - \mu_l)) = 0$ für $i \neq l$, da die y_is unabhängig sind.

Wegen der Nichtlinearität des Gleichungssystems (2.4) in $\boldsymbol{\beta}$ kann jenes nur numerisch gelöst werden. Wir verwenden im Folgenden die Newton-Raphson Methode, welche der Iterationsvorschrift

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \left(\boldsymbol{\mathfrak{I}}^{(m-1)} \right)^{-1} \mathbf{U}^{(m-1)} \tag{2.6}$$

genügt. Dabei beschreibt $\boldsymbol{\beta}^{(m)}$ den Parametervektor im m-ten Iterationschritt, während die rechte Seite im $m-1$-ten Schritt ausgewertet und $\mathbf{U}$ der Vektor mit Elementen U_j ist. Wird (2.6) auf beiden Seiten mit $\boldsymbol{\mathfrak{I}}^{(m-1)}$ multipliziert, so folgt

$$\boldsymbol{\mathfrak{I}}^{(m-1)} \boldsymbol{\beta}^{(m)} = \boldsymbol{\mathfrak{I}}^{(m-1)} \boldsymbol{\beta}^{(m-1)} + \mathbf{U}^{(m-1)}. \tag{2.7}$$

Wegen (2.5) können wir $\boldsymbol{\mathfrak{I}}^{(m-1)}$ als

$$\boldsymbol{\mathfrak{I}}^{(m-1)} = \frac{1}{\phi} \mathbf{X}^t \mathbf{W}^{(m-1)} \mathbf{X}$$

schreiben, wobei $\mathbf{W}^{(m)}$ eine $n \times n$ Diagonalmatrix mit Elementen

$$w_{ii}^{(m)} = \frac{1}{a_i V(\mu_i^{(m)})} \left(\frac{\partial \mu_i^{(m)}}{\partial \eta_i^{(m)}} \right)^2$$

ist. Der Ausdruck auf der rechten Seite von (2.7) ist der Vektor mit Elementen

$$\sum_{k=0}^{p-1}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{a_i\phi V(\mu_i^{(m-1)})}\left(\frac{\partial\mu_i^{(m-1)}}{\partial\eta_i^{(m-1)}}\right)^2\boldsymbol{\beta}_k^{(m-1)}+\sum_{i=1}^{n}\frac{(y_i-\mu_i^{(m-1)})x_{ij}}{a_i\phi V(\mu_i^{(m-1)})}\left(\frac{\partial\mu_i^{(m-1)}}{\partial\eta_i^{(m-1)}}\right),$$

welche im Ergebnis des $m-1$-ten Iterationsschritts evaluiert werden. Mit dem Vektor $\mathbf{z}^{(m)}=\left(z_1^{(m)},\ldots,z_n^{(m)}\right)^t$, welche die Elemente

$$z_i^{(m)}=\sum_{k=0}^{p-1}x_{ik}\boldsymbol{\beta}_k^{(m)}+(y_i-\mu_i^{(m)})\left(\frac{\partial\eta_i^{(m)}}{\partial\mu_i^{(m)}}\right)$$

besitzt, ist dies äquivalent zu

$$\frac{1}{\phi}\mathbf{X}^t\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}.$$

Also kann die Iterationsvorschrift (2.6) in die sogenannte Iteratively Reweighted Least Squares (IRLS) Notation umgeschrieben werden:

$$\mathbf{X}^t\mathbf{W}^{(m-1)}\mathbf{X}\boldsymbol{\beta}^{(m)}=\mathbf{X}^t\mathbf{W}^{(m-1)}\mathbf{z}^{(m-1)}. \tag{2.8}$$

Um den MLE $\hat{\boldsymbol{\beta}}$ zu berechnen haben die meisten statistischen Programme einen effizienten Algorithmus, welcher auf (2.8) basiert. Der Algorithmus stoppt, falls die Differenz zwischen $\boldsymbol{\beta}^{(m-1)}$ und $\boldsymbol{\beta}^{(m)}$ signifikant klein ist und wählt $\boldsymbol{\beta}^{(m)}$ als Maximum-Likelihood Schätzer.

2.2.4 Güte der Modellanpassung

Im Allgemeinen stimmen die fitted values $\hat{\mu}_i$ mit den y_i nicht überein und es stellt sich die Frage, wie groß die auftretenden Abweichungen sind. Denn während eine kleine Diskrepanz akzeptabel sein könnte, so ist eine große nicht vertretbar. Wir sind nun an einem Maß für diese Diskrepanz, auch Anpassungsgüte genannt, interessiert.

Seien n Beobachtungen gegeben, dann können die möglichen Modelle bis zu n Parameter aufweisen. Das einfachste Modell, das Nullmodell, hat nur einen Parameter, welcher die Erwartungswerte sämtlicher y_is repräsentiert. Dieses Modell ist meist zu einfach. Das andere Extrem, das volle (satuierte) Modell, hat n Parameter, einen für jede Beobachtung, und die fitted values entsprechen den Responsevariablen, also $\hat{\mu}_i=y_i$ für $i=1,\ldots,n$. Das volle Modell ist nicht sehr informativ, da es die Daten nicht zusammenfasst, jedoch gibt es uns einen Anhaltspunkt für die Messung der Diskrepanz für ein Modell mit p Parametern.

Betrachten wir nun ein Modell mit p Parameter und nehmen wir an, dass dieses korrekt ist, dann kann mittels der Likelihood-Quotienten Teststatistik zum Hypothesentest $H_0:\boldsymbol{\mu}=g^{-1}(\boldsymbol{\eta})$ gegen $H_1:\boldsymbol{\mu}\neq g^{-1}(\boldsymbol{\eta})$ die Anpassungsgüte beurteilt werden. Für den Likelihood-Quotienten folgt

$$\Lambda(\mathbf{y})=\frac{\sup\limits_{\boldsymbol{\mu}=g^{-1}(\boldsymbol{\eta})}L(\boldsymbol{\mu}|\mathbf{y})}{\sup\limits_{\boldsymbol{\mu}}L(\boldsymbol{\mu}|\mathbf{y})},$$

wobei der Zähler das Maximum der Likelihoodfunktion des betrachteten Modells und der Nenner das uneingeschränkte Maximum der Likelihoodfunktion beschreibt. Sei $\hat{\boldsymbol{\mu}}$ der MLE unter H_0; beim uneingeschränkten Modell wird die Likelihoodfunktion maximal für $\hat{\mu}_i = y_i$, $i = 1, \ldots, n$, also ergibt sich für den Quotienten

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\boldsymbol{\mu}}|\mathbf{y})}{L(\mathbf{y}|\mathbf{y})}.$$

Da unter gewissen Regularitätsbedingungen $-2\log(\Lambda(\mathbf{y}))$ asymptotisch einer Chi-Quadrat Verteilung genügt (siehe Dobson, 2001), wird $-2\log(\Lambda(\mathbf{y}))$ für die Bewertung der Güte der Modellanpassung verwendet und wir erhalten die skalierte Deviance

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\left(l(\hat{\boldsymbol{\mu}}|\mathbf{y}) - l(\mathbf{y}|\mathbf{y})\right),$$

wobei $D(\mathbf{y}, \hat{\boldsymbol{\mu}})$ Deviance genannt wird und $l(\boldsymbol{\mu}|\mathbf{y})$ die logarithmierte Likelihoodfunktion der Stichprobe beschreibt. Große Werte der skalierten Deviance weisen darauf hin, dass unser betrachtetes Modell eine schlechte Beschreibung der Daten bezüglich des satuierten Modells ist. Da $l(\mathbf{y}|\mathbf{y})$ unabhängig vom Modell und ein fester Wert, gegeben die Responses, ist, minimiert jener MLE $\hat{\boldsymbol{\mu}}$ die Deviance, welcher $l(\boldsymbol{\mu}|\mathbf{y})$ maximiert.

Ein andere wichtige Anpassungsgüte ist die generalisierte Pearson Statistik X^2, welche folgende Form besitzt

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

2.2.5 Residuen

Der nachstehende Abschnitt widmet sich den wichtigsten Residuen in GLMs und ist an Wood (2006) angelehnt.

Residuen spielen in der Klassischen Linearen Regression bei der Überprüfung der Modellanpassung eine wichtige Rolle, da sie alle Informationen der Daten enthalten, welche nicht vom Modell wiedergegeben werden. Bei den Generalisierten Linearen Modellen wird eine erweiterte Definition von Residuen notwendig, die für alle Verteilungen geeignet ist, welche die Normalverteilung ersetzen könnten.

Die naheliegendste Möglichkeit um die Residuen zu standardisieren ist, sie mit einer Größe zu skalieren, welche proportional zu ihrer Standardabweichung ist. Dies führt zu den Pearson Residuen

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

Werden nun diese gegen die fitted values geplottet, so sollten keine Trends im Erwartungswert und in der Varianz sichtbar sein. Der Name „Pearson Residuen“ ruht auf der Tatsache, dass die Summe der quadrierten Pearson Residuen die generalisierte Pearson Statistik X^2 ergibt, d.h.

$$\sum_{i=1}^{n} \left(r_i^P\right)^2 = X^2.$$

Da die Verteilung der Pearson Residuen in der Praxis asymmetrisch um Null sein kann, unterscheidet sich ihr Verhalten zu den Residuen der Klassischen Linearen Regression mehr als erwartet. In dieser Hinsicht sind die Deviance Residuen zu bevorzugen. Diese sind definiert als

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

wobei d_i die i-te Komponente der Deviance bezeichnet, also

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i.$$

Wird die Deviance für ein Modell berechnet, in welchem alle Parameter bekannt sind, so folgt die skalierte Deviance einer Chi-Quadrat Verteilung mit n Freiheitsgraden und dies suggeriert, dass $d_i \sim \chi_1^2$ und $r_i^D \sim N(0,1)$. Natürlich kann $D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) \sim \chi_n^2$ nicht in vernünftiger Weise auf ein einziges d_i angewendet werden, jedoch weist dies darauf hin, dass sich die Deviance Residuen bei einem Modell mit guter Anpassungsgüte ungefähr so wie $N(0,1)$-verteilte Zufallsvariablen verhalten.

2.2.6 Parametertests

In diesem Abschnitt versuchen wir die Parameter mittels verschiedener Konzepte zu testen. Ein Ansatz erfolgt über nested models, bei welchem folgende Hypothesen betrachtet werden:

$$H_0 : \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{q-1} x_{iq-1}$$
$$H_1 : \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{q-1} x_{iq-1} + \beta_q x_{iq} + \cdots + \beta_{p-1} x_{ip-1},$$

für $i = 1, \ldots, n$ und $q < p$, oder äquivalent

$$H_0 : \beta_q = \cdots = \beta_{p-1} = 0$$
$$H_1 : \beta_0, \ldots, \beta_{p-1} \text{ beliebig.}$$

Wir bezeichnen mit M_0 das unter H_0 betrachtete Modell und mit M jenes unter H_1. Also ist M_0 ein Untermodell von M.

Um H_0 gegen H_1 zu testen, können wir die Differenz der Deviance der beiden Modelle verwenden

$$D(M_0|M) = D(M_0) - D(M) = D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\phi \left(l(\hat{\boldsymbol{\mu}}_0|\mathbf{y}) - l(\hat{\boldsymbol{\mu}}|\mathbf{y}) \right),$$

wobei $\hat{\boldsymbol{\mu}}_0$ und $\hat{\boldsymbol{\mu}}$ die geschätzten Erwartungswerte unter dem jeweiligen Modell beschreiben. Man erkennt sofort, dass $D(M_0|M)/\phi$ mit der Likelihood-Quotienten Teststatistik übereinstimmt. Da $D(M_0)/\phi \sim \chi_{n-q}^2$ und $D(M)/\phi \sim \chi_{n-p}^2$ gilt, folgt $D(M_0|M)/\phi \sim \chi_{p-q}^2$. Also erhalten wir, falls der Dispersionsparamter ϕ bekannt ist, die Teststatistik

$$\frac{D(M_0) - D(M)}{\phi} = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\phi} \overset{H_0}{\sim} \chi_{p-q}^2.$$

Befindet sich der Wert der Testgröße im kritischen Bereich, so verwerfen wir die Nullhypothese zugunsten der Alternativhypothese mit der Begründung, dass Modell M eine signifikant bessere Beschreibung der Daten bietet. Im Fall, dass ϕ unbekannt ist, muss ϕ geschätzt werden und man zieht die folgende Teststatistik heran

$$\frac{\left(D(\mathbf{y}, \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}, \hat{\boldsymbol{\mu}})\right) / (p-q)}{D(\mathbf{y}, \hat{\boldsymbol{\mu}})/(n-p)} \overset{H_0}{\sim} F_{p-q,n-p}.$$

Ein weiterer Ansatz wird durch den sogenannten Wald Test repräsentiert. Wir betrachten dafür die ersten zwei Terme der Taylorreihe der Scorefunktion des Parametervektors $\mathbf{b}$ an der Entwicklungsstelle $\boldsymbol{\beta}$

$$\mathbf{U}(\mathbf{b}) = \mathbf{U}(\boldsymbol{\beta}) + \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{U}(\boldsymbol{\beta}) \cdot (\mathbf{b} - \boldsymbol{\beta}),$$

wobei $\mathbf{U}(\mathbf{b})$ dem Vektor der Scorefunktionen U_j an der Stelle $\boldsymbol{\beta} = \mathbf{b}$ entspricht. Approximieren wir $\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{U}(\boldsymbol{\beta})$ durch $\mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{U}(\boldsymbol{\beta})\right) = -\mathbb{E}\left(\mathbf{U}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta})^t\right) = -\boldsymbol{\mathfrak{I}}(\boldsymbol{\beta})$, so folgt

$$\mathbf{U}(\mathbf{b}) = \mathbf{U}(\boldsymbol{\beta}) - \boldsymbol{\mathfrak{I}}(\boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta}).$$

Setzen wir nun $\mathbf{b} = \hat{\boldsymbol{\beta}}$, dann gilt $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ und somit

$$\mathbf{U}(\boldsymbol{\beta}) = \boldsymbol{\mathfrak{I}}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

oder äquivalent

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{\mathfrak{I}}^{-1}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta}),$$

sofern $\boldsymbol{\mathfrak{I}}$ nichtsingulär ist. Betrachten wir $\boldsymbol{\mathfrak{I}}$ als eine Konstante, dann gilt $\mathbb{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{0}$, da der Erwartungswert der Scorefunktion Null ist, d.h. $\mathbb{E}(\mathbf{U}(\boldsymbol{\beta})) = \mathbf{0}$. Demzufolge ist $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, zumindest asymptotisch, sodass $\hat{\boldsymbol{\beta}}$ ein unverzerrter Schätzer von $\boldsymbol{\beta}$ ist. Für die Varianz-Kovarianz-Matrix von $\hat{\boldsymbol{\beta}}$ ergibt sich

$$\mathbb{E}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t\right) = \boldsymbol{\mathfrak{I}}^{-1}(\boldsymbol{\beta})\mathbb{E}\left[\mathbf{U}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta})^t\right]\left(\boldsymbol{\mathfrak{I}}^{-1}(\boldsymbol{\beta})\right)^t = \boldsymbol{\mathfrak{I}}^{-1}(\boldsymbol{\beta}),$$

weil $\boldsymbol{\mathfrak{I}} = \mathbb{E}\left(\mathbf{U}\mathbf{U}^t\right)$ und $\left(\boldsymbol{\mathfrak{I}}^{-1}\right)^t = \boldsymbol{\mathfrak{I}}^{-1}$, da $\boldsymbol{\mathfrak{I}}$ symmetrisch. Laut Fahrmeir und Kaufmann (1985) gilt sogar, dass $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{D} \text{Normal}(\mathbf{0}, n\boldsymbol{\mathfrak{I}}^{-1}(\boldsymbol{\beta}))$. Somit erhalten wir die Teststatistik

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \left(\boldsymbol{\mathfrak{I}}^{-1}(\hat{\boldsymbol{\beta}})\right)^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t \boldsymbol{\mathfrak{I}}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p,$$

die sogenannte Wald-Statistik.

2.3 Logistische Regression

Nehmen wir nun an, dass jede der Responsevariablen Y_i nur zwei Werte, der Einfachheit halber mit 0 und 1 bezeichnet, annehmen kann. Beobachtungen dieser Art findet man in vielen Bereichen, z.B. in klinischen Studien, in welcher die Patienten nach Ablauf der Studie entweder genesen oder nicht genesen sind. Wir definieren also die binäre Zufallsvariable

$$Y_i = \begin{cases} 1, & \text{falls Ereignis eintritt} \\ 0, & \text{falls Ereignis nicht eintritt} \end{cases}$$

mit Wahrscheinlichkeiten $\mathbb{P}(Y_i = 1) = \pi_i$ und $\mathbb{P}(Y_i = 0) = 1 - \pi_i$.

Oft ist es in der Praxis klug die Daten mit binären Responses, welche dieselben Prädikorvektor besitzen, zusammenzufassen und als eine Klasse zu gruppieren. Wir nehmen also an, dass für die i'-te Ausprägungsmöglichkeit, charakterisiert durch den Prädiktorvektor $(1, x_{i'1}, \ldots, x_{i'p-1})^t$, $m_{i'}$ Beobachtungen vorliegen. Dies bedeutet mit anderen Worten, dass von der Stichprobe, $n = \sum_{i'=1}^{k} m_{i'}$, $m_{i'}$ den Vektor $(1, x_{i'1}, \ldots, x_{i'p-1})^t$ teilen und somit die Klasse i' bilden. Für diese Klassen wird nun die neue Responsevariable $m_{i'} y^*_{i'}$ eingeführt, wobei $y^*_{i'}$ die relative Häufigkeit der Erfolge in der Klasse i' wiedergibt, mit $i' = 1, \ldots, k$. Demnach beschreibt $m_{i'} y^*_{i'}$ die Anzahl der Erfolge in der Klasse i' und genügt der Binomialverteilung. Zur Illustrierung der alternativen Darstellungsmöglichkeit der Daten wird in der Tabelle 2.3 ein Beispiel angegeben. Auf der linken Seite der Tabelle sind die Daten nach Beobachtungen i gelistet, während auf der rechten Seite die Daten zu Klassen der Größe $m_{i'}$ zusammengefasst werden.

Daten in gewöhnlicher Form gelistet			Daten in Klassen gelistet		
i	(x_{i1}, x_{i2})	y_i	$(x_{i'1}, x_{i'2})$	$m_{i'}$	$m_{i'} y^*_{i'}$
1	1, 1	0	1, 1	2	1
2	1, 2	1	1, 2	3	2
3	1, 2	0	2, 1	1	0
4	2, 1	0	2, 2	1	1
5	2, 2	1			
6	1, 2	1			
7	1, 1	1			

Tabelle 2.3: Alternative Möglichkeiten um dieselben Daten zu präsentieren (McCullagh und Nelder, 1989).

Falls jede Beobachtung eine unterschiedliche Prädiktorvariable besitzt, dann gilt $k = n$ sowie $m_1 = \cdots = m_n = 1$ und die Responsevariable ist binär. Die ungruppierten Daten können somit als ein Spezialfall angesehen werden, sodass wir ausschließlich gruppierte Daten betrachten. Um unsere in den bisherigen Kapiteln verwendete Notation aufrechtzuerhalten, setzen wir $i = i'$, $y_i = y_{i'}$ bzw. $n = k$ und wir erhalten das folgende Modell:

$$m_i y_i \overset{ind}{\sim} \text{Binomial}(m_i, \pi_i) \qquad i = 1, \ldots, n.$$

Nun werden einige wichtige Eigenschaften für das obige Modell berechnet.

2.3.1 Erwartungswert und Varianz von y_i

Da die Wahrscheinlichkeitsfunktion der relativen Häufigkeiten, $y_i = 0, 1/m_i, 2/m_i, \ldots, 1$, in die Form

$$\begin{aligned} f(y_i|m_i, \pi_i) = \mathbb{P}(Y_i = y_i) = \mathbb{P}(m_i Y_i = m_i y_i) &= \binom{m_i}{m_i y_i} \pi_i^{m_i y_i} (1-\pi_i)^{m_i - m_i y_i} \\ &= \exp\left(\log \binom{m_i}{m_i y_i} + m_i y_i \log \pi_i + m_i (1-y_i) \log(1-\pi_i) \right) \\ &= \exp\left(\frac{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) - \log\left(\frac{1}{1-\pi_i}\right)}{1/m_i} + \log \binom{m_i}{m_i y_i} \right) \end{aligned}$$

mit

$$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right), \qquad \phi = 1$$

und

$$a_i = \frac{1}{m_i}, \qquad b(\theta_i) = \log\left(\frac{1}{1-\pi_i}\right) = \log\left(1 + \exp(\theta_i)\right), \qquad c(y_i, \phi) = \log \binom{m_i}{m_i y_i}$$

umgeschrieben werden kann, ist die Verteilung von relativen Häufigkeiten ein Mitglied der linearen Exponentialfamilie (siehe auch Tabelle 2.2). Mit (2.3) erhalten wir die ersten beiden Momente

$$\begin{aligned} \mathbb{E}(y_i) = \mu_i = b'(\theta_i) &= \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = \pi_i, \\ \mathrm{Var}(y_i) = a_i \phi b''(\theta_i) &= \frac{1}{m_i} \left(\frac{\exp(\theta_i)}{1+\exp(\theta_i)} \right)' \\ &= \frac{1}{m_i} \frac{\exp(\theta_i)}{\left(1+\exp(\theta_i)\right)^2} = \frac{1}{m_i} \pi_i (1-\pi_i). \end{aligned}$$

2.3.2 Linkfunktion

Wie bereits im vorigen Abschnitt beschrieben, verbindet die Linkfunktion

$$g(\mu_i) = \eta_i = \sum_{j=0}^{p-1} x_{ij} \beta_j, \qquad i = 1, \ldots, n$$

den linearen Prädiktor mit dem Erwartungswert. Da bei der standardisierten Binomialverteilung $0 < \mu_i = \pi_i < 1$ gilt, muss also $g(\pi_i)$ das Einheitsintervall auf die reelle Zahlengerade abbilden. Dafür steht eine große Auswahl an Linkfunktionen zur Verfügung. Die folgenden drei Funktionen werden üblicherweise verwendet.

- Logitlink (die kanonische Linkfunktion):

$$g_1(\pi) = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{m\pi}{m-m\pi}\right) = \theta = \eta$$

  bzw.

$$g_1^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} = \pi.$$

- Probitlink:

$$g_2(\pi) = \text{probit}(\pi) = \Phi^{-1}(\pi) = \eta \quad \text{bzw.} \quad g_2^{-1}(\eta) = \Phi(\eta) = \pi.$$

- Komplementäre log-log-Link:

$$g_3(\pi) = \log(-\log(1-\pi)) = \eta \quad \text{bzw.} \quad g_3^{-1}(\eta) = 1 - \exp(-\exp(\eta)) = \pi.$$

Eine weitere Möglichkeit ist der log-log-Link

$$g_4(\pi) = -\log(-\log(\pi)) = \eta \quad \text{bzw.} \quad g_4^{-1}(\eta) = \exp(-\exp(-\eta)) = \pi,$$

welcher jedoch selten benutzt wird, da sein Verhalten für $\pi < 1/2$ inadäquat ist (McCullagh und Nelder, 1989).

Nun sind wir am Vergleich der vier Funktionen interessiert. Eine graphische Gegenüberstellung dieser wird in Abbildung 2.2 angezeigt. Man erkennt, dass der Logitlink und der Probitlink symmetrische Linkfunktionen sind, während die anderen beiden Funktionen asymmetrisch sind. Für die künftigen Analysen und Berechnungen werden wir meist den Logitlink verwenden, nicht nur wegen seiner einfacheren theoretischen Eigenschaften, sondern hauptsächlich wegen seiner einfachen Interpretation.

2.3.3 Interpretation der Parameter

Wir betrachten nun ungruppierte Daten der Form

$$y_i \overset{ind}{\sim} \text{Binomial}(1, \pi_i) \qquad i = 1, \ldots, n.$$

Dann kann der Logitlink auch als der Logarithmus der Odds (Chance oder Quote) geschrieben werden:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{\mathbb{P}(y_i=1)}{1-\mathbb{P}(y_i=1)}\right) = \log(\text{odds}_i) = \sum_{j=0}^{p-1} x_{ij}\beta_j = \eta_i.$$

Odds geben die Quote für das Eintreten eines Ereignisses im Verhältnis zu dem Nicht-Eintreten des Ereignisses wieder. Diese können nun einfach aus der obigen Gleichung berechnet werden:

$$\text{odds}_i = \frac{\mathbb{P}(y_i=1)}{1-\mathbb{P}(y_i=1)} = \exp\left(\sum_{j=0}^{p-1} x_{ij}\beta_j\right) = \prod_{j=0}^{p-1} \exp\left(x_{ij}\beta_j\right) = \prod_{j=0}^{p-1} \exp\left(\beta_j\right)^{x_{ij}}.$$

Dieses Ergebnis lässt sich folgenderweise interpretieren: Erhöhen wir z.B. den Prädiktor x_{i2} um eine Einheit, während die anderen Prädiktoren festgehalten werden, so wächst die Eintrittsquote von $y_i = 1$ multiplikativ um $\exp(\beta_2)$.

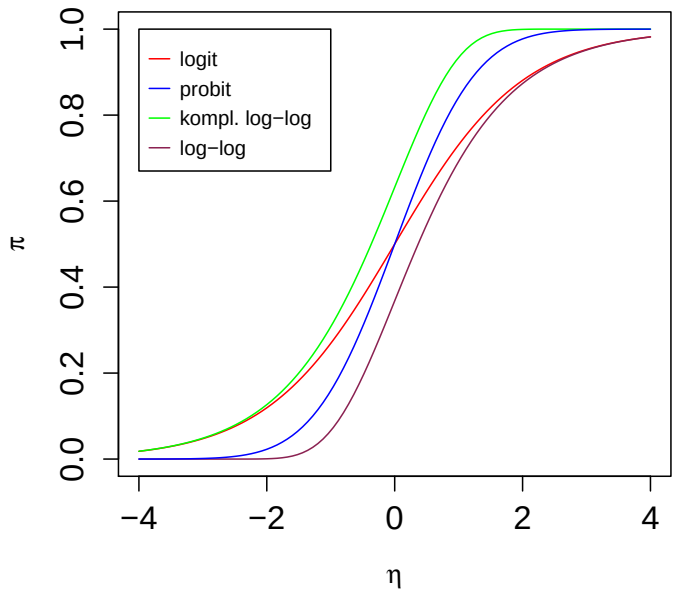

Abbildung 2.2: Abbildung des linearen Prädiktors auf π durch die vier verschiedenen Linkfunktionen.

2.3.4 Log-Likelihoodfunktion und Deviance

Hier sei noch kurz die Berechnung der Log-Likelihoodfunktion und der Deviance für die logistische Regression erwähnt. Als Log-Likelihoodfunktion erhalten wir

$$\begin{aligned} l(\boldsymbol{\pi}|\mathbf{y}) = & \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a_i\phi} + c(y_i, \phi) \right) \\ & \sum_{i=1}^{n} \left(m_i y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) - m_i \log\left(\frac{1}{1-\pi_i}\right) + \log\binom{m_i}{m_i y_i} \right), \end{aligned}$$

und für die Deviance folgt

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\pi}}) = & -2\left(l(\hat{\boldsymbol{\pi}}|\mathbf{y}) - l(\mathbf{y}|\mathbf{y})\right) \\ = & -2\sum_{i=1}^{n} \left[m_i y_i \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) - m_i \log\left(\frac{1}{1-\hat{\pi}_i}\right) + \log\binom{m_i}{m_i y_i} \right. \\ & \left. - m_i y_i \log\left(\frac{y_i}{1-y_i}\right) + m_i \log\left(\frac{1}{1-y_i}\right) - \log\binom{m_i}{m_i y_i} \right] \\ = & 2\sum_{i=1}^{n} m_i \left((1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right) + y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) \right). \end{aligned}$$

2.3.5 Überdispersion

Unter der Bezeichnung Überdispersion verstehen wir, dass die Varianz der Response größer als die nominale Varianz unter dem betrachteten Modell ist. Da wir nun absolute Häufigkeiten betrachten, meinen wir also unter dem Begriff eine Situation, in der die Responsevarianz die binomiale Varianz übersteigt. Überdispersion ist ein sehr häufiges Phänomen; laut McCullagh und Nelder (1989) ist sie in der Praxis sogar eher die Regel als die Ausnahme.

Es gibt eine Vielzahl von Gründen für das Auftreten von Überdispersion. Der einfachste und wahrscheinlich geläufigste Beweggrund für die Notwendigkeit eines Dispersionsparameters im Verteilungsmodells kann anhand von „Häufungen in der Population“ (Cluster) aufgezeigt werden. Familien, Haushalte und Nachbarschaften sind Beispiele für natürlich vorkommende Cluster in der Population. Wir nehmen nun an, dass in der i-ten Clusterumgebung m_i korrelierte, binäre Bernoulli-Variablen $y_{i1}, \ldots, y_{im_i}$ mit Erwartungswert $\mathbb{E}(y_{ij}) = \pi_i$ und Varianz $\mathrm{Var}(y_{ij}) = \pi_i(1-\pi_i)$ beobachtet werden, wobei in einem Cluster i alle möglichen Paare denselben Korrelationskoeffizienten besitzen, nämlich

$$\rho = \frac{\mathbb{E}(y_{ij}y_{ik}) - \mathbb{E}(y_{ij})\mathbb{E}(y_{ik})}{\sqrt{\mathrm{Var}(y_{ij})\,\mathrm{Var}(y_{ik})}} = \frac{\mathbb{P}(y_{ij}=1, y_{ik}=1) - \pi_i^2}{\pi_i(1-\pi_i)}, \qquad j \neq k;\ \ i = 1, \ldots, n.$$

Für die Summe $y_i = \sum_{j=1}^{m_i} y_{ij}$ der korrelierten Bernoulli-Variablen in Cluster i ergibt sich

$$\begin{aligned}
\mathbb{E}(y_i) &= m_i\pi_i \\
\mathrm{Var}(y_i) &= \sum_{j=1}^{m_i} \mathrm{Var}(y_{ij}) + \sum_{j\neq k}^{m_i} \mathrm{Cov}(y_{ij}, y_{ik}) \\
&= m_i\pi_i(1-\pi_i) + m_i(m_i-1)\pi_i(1-\pi_i)\rho \\
&= m_i\pi_i(1-\pi_i)(1+(m_i-1)\rho) = \phi_i m_i \pi_i(1-\pi_i).
\end{aligned}$$

Man erkennt sofort, dass sich die Varianz der Summe korrelierter Null-Eins-Variablen um den Term $\phi_i = (1+(m_i-1)\rho)$ von der Varianz einer Binomialverteilung unterscheidet. Besitzt ρ einen positiven Wert, so spricht man von einer Überdispersion. Dieser kann nur auftreten, falls $m_i > 1$. Ist nämlich $m_i = 1$, dann folgt $\phi_i = 1$ und die beiden Varianzen stimmen wieder überein.

Um mit dem Problem der Überdispersion umzugehen wird gelegentlich die Beta-Binomialverteilung verwendet. Für weitere Informationen siehe Williams (1982). Eine alternative Möglichkeit für das Lösen der aufgetretenen Problematik wird in McCullagh und Nelder (1989) vorgeschlagen. Dafür wird der Einfachheit halber vorausgesetzt, dass die Größen der Cluster gleich sind, also $m_i = m$ für $i = 1, \ldots, n$ gilt. Daraus folgt für den Dispersionsparameter, dass $\phi_i = \phi = (1+(m-1)\rho)$ für alle i, und der Effekt der Überdispersion besitzt jetzt die Form

$$\begin{aligned}
\mathbb{E}(y_i) &= m_i\pi_i \\
\mathrm{Var}(y_i) &= \phi m_i \pi_i(1-\pi_i),
\end{aligned}$$

d.h. der Erwartungswert bleibt unberührt, jedoch wird die Varianz um den unbekannten Faktor ϕ vergrößert. Werden nun die Methoden aus Kapitel 2 auf das logistische Modell mit geänderter Varianzstruktur angewendet, so wird der gleiche Schätzer für den Parameter $\boldsymbol{\beta}$ generiert, den wir für den Fall einer Binomialverteilung erhalten würden. Es entsteht auch keine Veränderung bei der Deviance. Der Unterschied, der jedoch auftritt, ist die veränderte Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$, denn für sie gilt nun

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathfrak{I}^{-1}(\hat{\boldsymbol{\beta}}) = \phi(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}.$$

Schlussendlich bleibt noch die einzig offene Frage, wie der unbekannte Dispersionsparameter geschätzt werden kann. Dafür empfehlen die beiden Autoren den folgenden Schätzer:

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

wobei X^2 die generalisierte Pearson Statistik ist.

Werden keine gleichen Clustergrößen angenommen, so ändert sich $\hat{\boldsymbol{\beta}}$ und wegen den unterschiedlichen Dispersionsparametern ϕ_i muss auch ein expliziter Schätzer für ρ gefunden werden. Genauere Details können in Williams (1982) nachgelesen werden.

2.3.6 Beispiel

Zur Veranschaulichung der bisher diskutierten Modelle wird ein sehr einfaches Beispiel betrachtet. Der verwendete Datensatz `menarche` stammt aus der `MASS`-Bibliothek von Venables und Ripley (2002) und beinhaltet Informationen einer Studie von Milicer und Szczotka (1966), in welcher der Anteil der Mädchen verschiedener Altersgruppen gemessen wurde, für welche die Menarche bereits eingetreten war. Die Datei beinhaltet 25 Datenpunkte mit den folgenden drei Informationen: „Age“ beschreibt das Durchschnittsalter der jeweiligen Altersgruppe, „Total“ ist die Gesamtanzahl der Mädchen in jeder Gruppe und „Menarche“ gibt die Zahl der Mädchen einer Altersgruppe wieder, welche die Menarche erreicht hatten. Die beobachteten Daten befinden sich in Tabelle 2.4.

Wir wollen im Folgenden anhand des Alters der Mädchen die Wahrscheinlichkeit des Eintritts der Menarche schätzen. Dafür wird ein Modell konstruiert, welches versucht, den Anteil der jungen Frauen, welche die Menarche schon erreicht haben, zu erfassen:

$$m_i y_i \sim \mathrm{Binomial}(m_i, \pi_i),$$

wobei m_i die Anzahl der Mädchen im Alter x_i angibt und y_i die Proportion der Mädchen mit bereits eingetretener Menarche für das jeweilige Alter x_i bezeichnet. Mit dem Logitlink

$$\mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

folgt für die Wahrscheinlichkeit, dass die Menarche schon eingesetzt hat, bzw. für den erwarteten Anteil

$$\pi_i = \mathbb{E}(y_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Age	Total	Menarche
9.21	376	0
10.21	200	0
10.58	93	0
10.83	120	2
11.08	90	2
11.33	88	5
11.58	105	10
11.83	111	17
12.08	100	16
12.33	93	29
12.58	100	39
12.83	108	51
13.08	99	47
13.33	106	67
13.58	105	81
13.83	117	88
14.08	98	79
14.33	97	90
14.58	120	113
14.83	102	95
15.08	122	117
15.33	111	107
15.58	94	92
15.83	114	112
17.58	1049	1049

Tabelle 2.4: Daten einer Studie über den Eintrittszeitpunkt der Menarche.

Es macht Sinn, zuerst die beobachteten Anteile gegen die Altersgruppen zu plotten.

```
> library(MASS)
> attach(menarche)

> y <- Menarche/Total
> plot(Age, y, xlab="Alter", ylab="Menarche/Total", cex.lab=0.8)
```

Der resultierende Graph wird in Abbildung 2.3 dargestellt. Man erkennt sofort, dass laut unseren Beobachtungen die Wahrscheinlichkeit für das Eintreten der Menarche im Alter von 12 und 14 massiv steigt, während von 14 bis 16 nur mehr ein leichter Anstieg zu beobachten ist.

Es wird nun versucht die unbekannten Parameter des obigen Modells mittels der `glm()`-Funktion zu schätzen. Als Response wird eine Matrix mit 2 Spalten angelegt, in welcher die erste Spalte die jeweilige Anzahl der jungen Frauen nach der Menarche und die zweite die Anzahl der jungen Frauen vor der Menarche beschreibt. Durch das „∼“-Zeichen wird

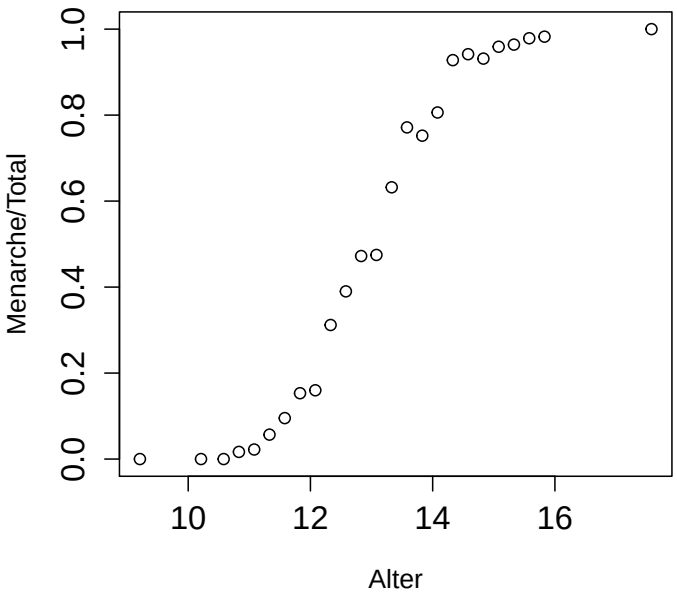

Abbildung 2.3: Beobachtete Proportionen der Mädchen mit bereits eingesetzter Menarche zum jeweiligen Alter.

die Responsevariable mit der erklärenden Variable `Age` verbunden. Anschließend wird die Verteilung und die Linkfunktion übergeben, zu guter Letzt der Name des Datensatzes. Wir erhalten also

```
> mod.1 <- glm(cbind(Menarche,Total-Menarche)~Age, family=binomial(link=logit),
+              data=menarche)
> mod.1

Call:  glm(formula = cbind(Menarche, Total - Menarche) ~ Age,
           family = binomial(logit), data = menarche)

Coefficients:
(Intercept)          Age
    -21.226        1.632

Degrees of Freedom: 24 Total (i.e. Null);  23 Residual
Null Deviance:     3694
Residual Deviance: 26.7  AIC: 114.8
```

Die `Null Deviance` gibt die Deviance des Modells mit lediglich dem konstanten Term an, während die `Residual Deviance` die Deviance des gefitteten Modells bezeichnet. Man erkennt, dass die letztere mit 26.7 für eine $\chi^2_{23}-$Zufallsvariable plausibel zu sein scheint,

denn die Wahrscheinlichkeit, dass eine $\chi^2_{23}-$Variable größer als 26.7 ist, beträgt

```
> 1-pchisq(26.7,23)
[1] 0.2689471
```

Also unterscheiden sich die gefitteten Werte des Modells nicht signifikant von den beobachteten Werten. Mit `AIC` wird das Akaike Informationskriterum gekennzeichnet, welches ein Kriterium zur Auswahl eines Modells aus einer Klasse von GLMs ist. Je kleiner der Wert des AIC, welches als

$$\text{AIC} = 2\left(-l(\hat{\boldsymbol{\theta}}|\mathbf{y}) + p\right)$$

berechnet wird, desto besser ist das Modell. Dabei beschreibt p die Gesamtanzahl an schätzenden Parametern.

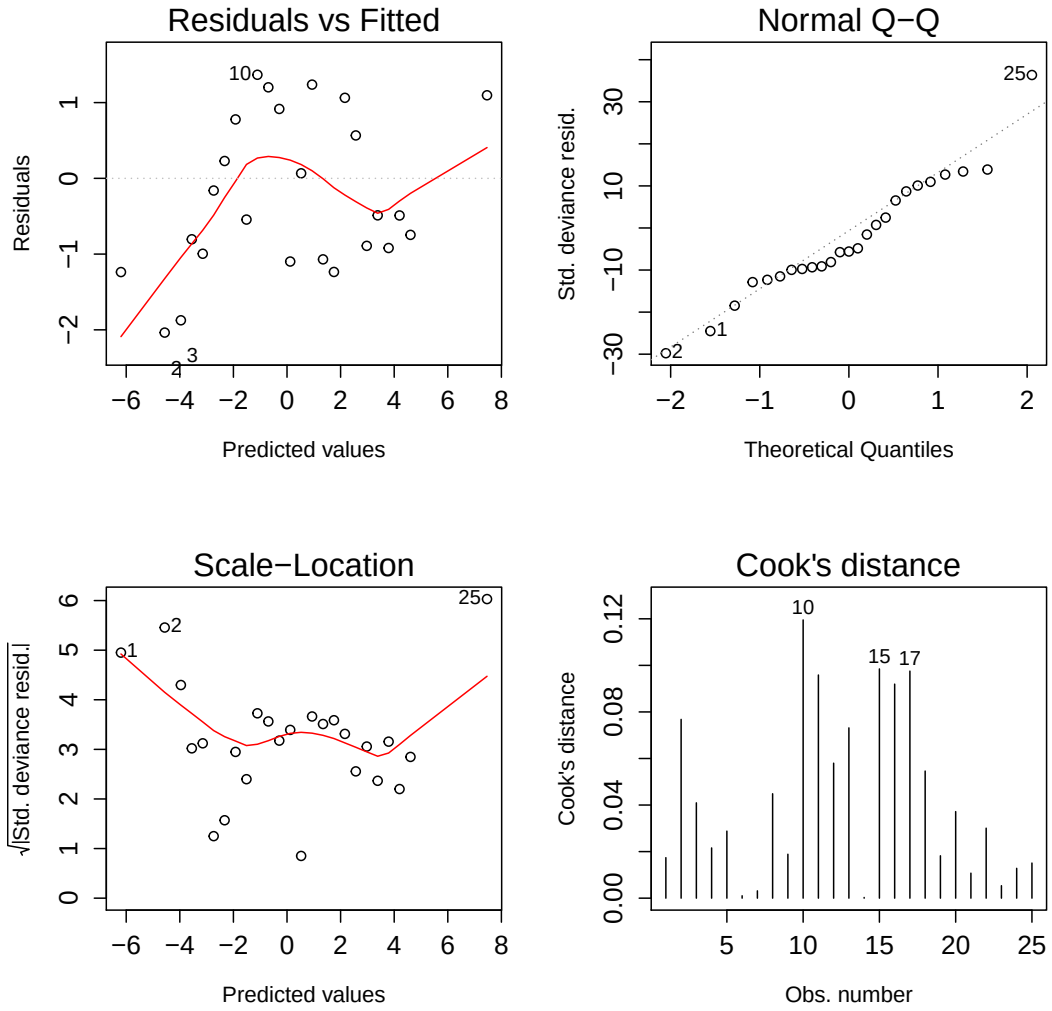

Abbildung 2.4: Diagonistische Plots zur Überprüfung von `mod.1`.

Nun werden die Residuen untersucht. Die Diagramme von

```
> par(mfrow=c(2,2))
> plot(mod.1, which=c(1:4), cex.lab=0.8)
```

werden in Abbildung 2.4 dargestellt. Im linken oberen Bild werden die Deviance Residuen gegen die `Predicted values` aufgetragen, wobei letztere hier nicht die gefitteten Werte der Response sind, sondern die des linearen Prädiktors. Das rechte obere Bild ist ein Normal-Quantil-Quantil-Plot. Hierfür werden die standardisierten Deviance Residuen sortiert und anschließend gegen die theoretischen Quantile der Normalverteilung geplottet. Kleine Abweichungen von der linearen Beziehung im QQ-Plot sind bei den GLMs keine Seltenheit. Hier muss noch kurz die Definition der standardisierten Deviance Residuen eingeschoben werden. Diese sind durch

$$r_i^{SD} = \frac{r_i^D}{\sqrt{\hat{\phi}\,(1 - \mathbf{H}_{ii})}}$$

festgelegt, wobei die Hatmatrix $\mathbf{H}$ die Form $\mathbf{H} = \hat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}^t\hat{\mathbf{W}}\mathbf{X}^t\right)^{-1}\mathbf{X}\hat{\mathbf{W}}^{1/2}$ besitzt (für genauere Informationen bzgl. der Hatmatrix siehe Pregibon, 1981). Ein sogenannter Scale Location Plot wird im linken unteren Bild gezeigt. Dieser ist ähnlich zum ersten Bild, jedoch wird diesmal anstelle der Deviance Residuen die Wurzel des Absolutbetrags der standardisierten Deviance Residuen verwendet. Das letzte Bild (rechts unten) gibt einen Cook's Distance Plot wieder, welcher angibt, ob im Datensatz einflussreiche Punkte vorhanden sind. Solche einflussreichen Datenpunkte sind Beobachtungen, ohne die das geschätzte Modell signifikant anders wäre. Bezeichnen wir mit $\hat{\boldsymbol{\beta}}^{[-i]}$ den geschätzten Parametervektor für das Modell, welches die i-te Beobachtung nicht berücksichtigt, und mit p die Anzahl der Parameter im linearen Prädiktor, dann wird die Cook-Distanz durch

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}}^{[-i]} - \hat{\boldsymbol{\beta}}\right)^t \left(\mathbf{X}^t\hat{\mathbf{W}}\mathbf{X}\right)\left(\hat{\boldsymbol{\beta}}^{[-i]} - \hat{\boldsymbol{\beta}}\right)\Big/ p}{\hat{\phi}}$$

definiert (siehe McCullagh und Nelder, 1989). Aufgrund der geringen Anzahl an Daten ist es sehr schwierig die obigen Diagramme zu interpretieren, jedoch scheint es keine Hinweise für eine fehlerhafte Modellierung zu geben. Möglicherweise kann ein leichter Trend im Erwartungswert der Residuen abgelesen werden.

Zu guter Letzt wird der Graph der gefitteten Werte im Bild der beobachteten Proportionen (siehe Abbildung 2.5) betrachtet. Man sieht leicht, dass mit unserem gewählten Modell eine sehr gute Anpassung möglich ist.

```
> plot(Age, y, xlab="Alter", ylab="Menarche/Total", cex.lab=0.8)
> lines(Age, fitted(mod.1))
```

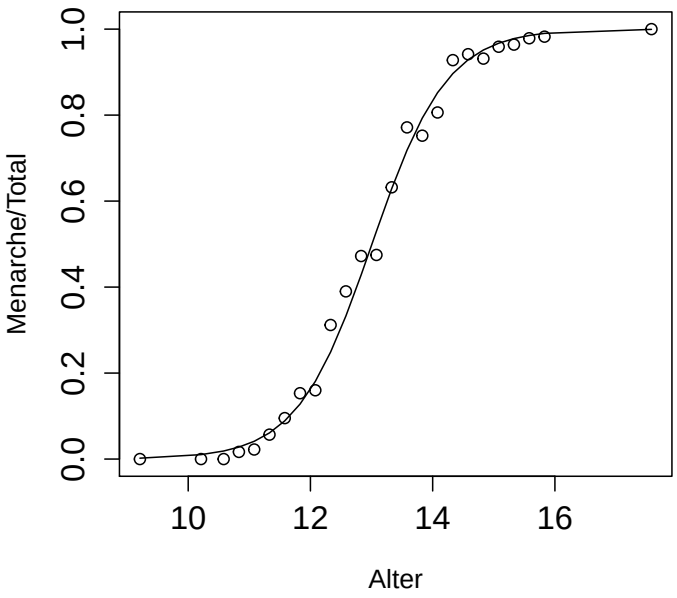

Abbildung 2.5: Beobachtete und geschätzte Wahrscheinlichkeit (`mod.1`) der bereits eingetretenen Menarche zum jeweiligen Alter.

3 Nichtparametrische Regression

3.1 Additive Modelle

Die Klasse der Generalisierten Additiven Modelle (GAM), eingeführt von Hastie und Tibshirani (1986), sind eine Erweiterung der GLMs, wobei der lineare Prädiktor nicht mehr nur rein parametrisch modelliert wird, sondern zusätzlich eine Summe von glatten Funktionen der Prädiktorvariablen enthält. Das Modell hat also beispielsweise die Form

$$g(\mu_i) = \mathbf{x}_i^{*\,t}\boldsymbol{\beta} + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}, x_{i4}) + \dots \tag{3.1}$$

mit

$$\mu_i = \mathbb{E}(y_i), \qquad y_i \overset{ind}{\sim} \text{Exponentialfamilie}(\theta_i).$$

Hierbei ist y_i die Responsevariable, $\mathbf{x}_i^*$ entspricht dem Prädiktorvektor für die parametrischen Modellkomponenten, $\boldsymbol{\beta}$ stellt den Parametervektor dar und $f_k(\cdot)$ ist eine glatte Funktion der Prädiktorvariablen. Eine Funktion wird als glatt bezeichnet, falls sie bis zur gewünschten Ordnung stetig differenzierbar ist.

Da GAMs aus Additiven Modellen (AMs) folgen wie GLMs aus Linearen Modellen, betrachten wir zur Einführung die Klasse der Additiven Modelle, wobei wir uns im Folgenden an Wood (2006) und Fahrmeir, Kneib und Lang (2009) orientieren. Weitere erwähnenswerte Quellen sind Buja, Hastie und Tibshirani (1989) und Ruppert et al. (2003). Der Einfachheit halber und um unnötige technische Details zu vermeiden, sei unser zu untersuchendes Modell das Folgende:

$$y_i = f(x_i) + \epsilon_i, \tag{3.2}$$

wobei y_i die Responsevariable, x_i die erklärende Variable, f eine glatte Funktion und ϵ_i eine i.i.d. $N(0, \sigma^2)$ Zufallsvariable ist. Zur weiteren Vereinfachung wird angenommen, dass x_i im Intervall $[0, 1]$ liegt. Es stellen sich nun zwei Fragen:

1. Wie kann die Funktion f geschätzt werden?
2. Wie glatt soll f sein?

3.1.1 Regression Splines

Zur Beantwortung der ersten Frage wird versucht (3.2) in ein Lineares Modell umzugestalten, um die Methoden der vorhergehenden Kapiteln anwenden zu können. Zu diesem Zweck wird angenommen, dass die Funktion f die Form

$$f(x) = \sum_{j=0}^{q-1} b_j(x)\beta_j \tag{3.3}$$

besitzt, wobei $b_j(x)$ für die j-te Basisfunktion und $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{q-1})^t$ für den unbekannten Parametervektor steht. Das Einsetzen von (3.3) in (3.2) ergibt eindeutig ein Lineares Modell.

Polynomiale Basis
Die Aufgabe besteht also darin, eine geeignete Basis für f zu finden. Wichtig ist hier zu erwähnen, dass eine Vielzahl von Basen existieren, im Folgenden jedoch nur zwei einfache untersucht werden. Die erste ist die polynomiale Basis. In diesem Fall wird angenommen, dass f ein Polynom der Ordnung $q-1$ ist. Somit folgen für die Basisfunktionen $b_j(x)$

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

und demzufolge

$$f(x) = \sum_{j=0}^{q-1} x^j \beta_j. \tag{3.4}$$

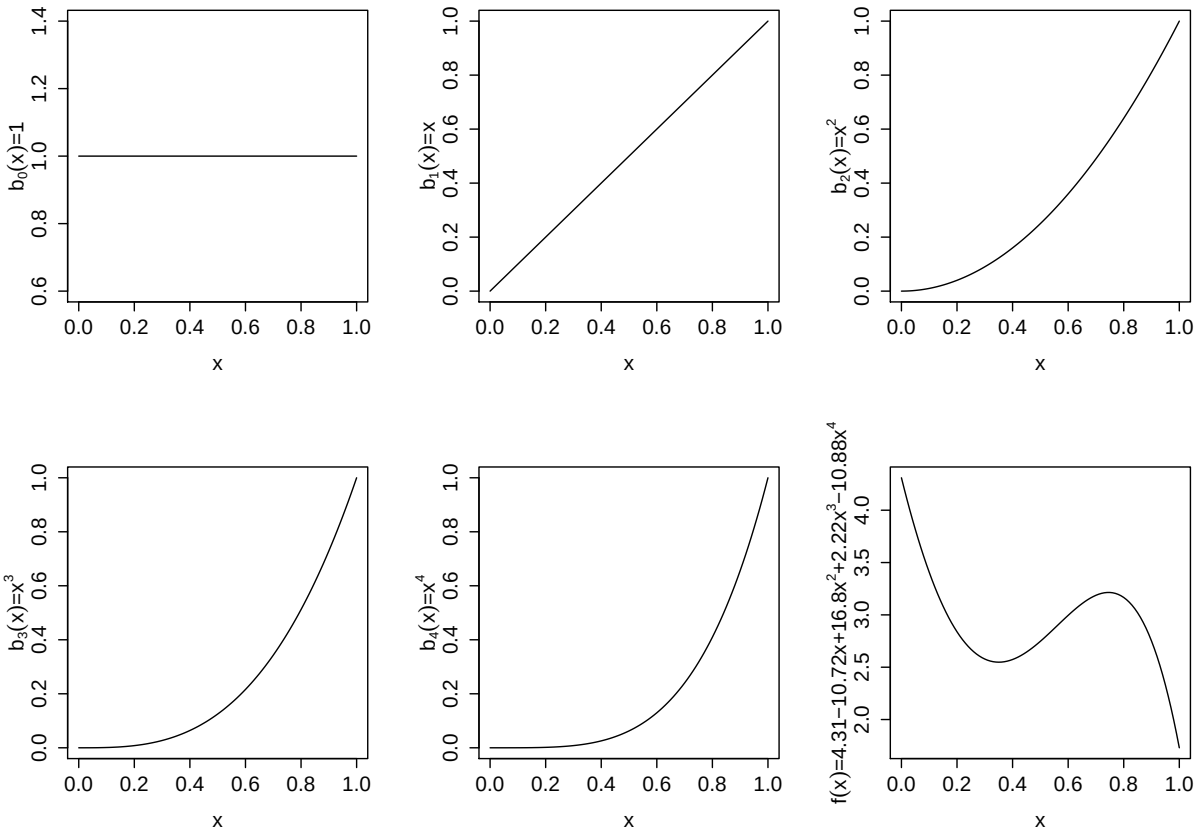

Abbildung 3.1: Die fünf Basisfunktionen eines Polynoms vierter Ordnung und ein mögliches resultierendes Polynom $f(x) = 4.31 - 10.72x + 16.8x^2 + 2.22x^3 - 10.88x^4$ (Wood, 2006).

Das Prinzip der polynomialen Basis wird in Abbildung 3.1 veranschaulicht. Die ersten fünf Bilder stellen die fünf Basisfunktionen eines Polynoms vierter Ordnung dar. Jede dieser

Basisfunktionen wird mit einem Parameter, β_j, multipliziert und anschließend summiert. Ein Beispiel für eine mögliche resultierende Funktion $f(x)$ wird im unteren rechten Graph angezeigt.

Um die Funktion f in (3.4) zu erhalten, müssen die Parameter β_j geschätzt werden, sowie die Anzahl der Polynome q oder äquivalent die maximale polynomiale Ordnung $q-1$. Bei gegebener Ordnung können die Parameter einfach durch Anwenden der Methoden der Linearen Regression geschätzt werden.

Da polynomiale Basisfunktionen für den ganzen Definitionsbereich (derzeit $[0,1]$) angepasst werden, besitzen sie hinsichtlich Flexibilität einige Schwächen. Um jene zu beseitigen werden sogenannte Splines verwendet. Diese sind Funktionen, die sich stückweise aus Polynomen zusammensetzen. Eine häufig verwendete ist die folgende:

Kubische Spline Basis
Sind die zusammengefügten Funktionen kubische Polynome, so spricht man von einem Kubischen Spline. An den Nahtstellen, an denen zwei Polynomstücke zusammenstoßen, auch Knoten genannt, wird gefordert, dass die Werte der beiden Polynome übereinstimmen, sowie auch die Werte der ersten und zweiten Ableitung. Die Position der Knoten muss gewählt werden; für gewöhnlich werden die Knoten auf den Bereich der beobachteten x-Werte gleichmäßig oder an den Quantilen der Verteilung von x platziert, und haben im Folgenden die Bezeichnung $\{x_i^* : i = 1, \ldots, q-2\}$.

Bei gegebenen Knoten findet man einige verschiedene, jedoch äquivalente Möglichkeiten vor um die Basen der Kubischen Splines niederzuschreiben. Eine laut Wood (2006) einfach zu verwendende Basis, deren Herleitungen z.B. in Gu (2002) nachgelesen werden können, wird nachstehend angeführt:

$$b_0(x) = 1, \qquad b_1(x) = x, \qquad b_{i+1}(x) = R(x, x_i^*) \quad \text{für } i = 1, \ldots, q-2,$$

mit

$$\begin{aligned} R(x,z) = &\left(\left(z-\frac{1}{2}\right)^2 - \frac{1}{12}\right)\left(\left(x-\frac{1}{2}\right)^2 - \frac{1}{12}\right)\frac{1}{4} \\ &- \left(\left(|x-z| - \frac{1}{2}\right)^4 - \frac{1}{2}\left(|x-z| - \frac{1}{2}\right)^2 + \frac{7}{240}\right)\frac{1}{24}. \end{aligned}$$

Für unser Modell (3.2) folgt bei Anwendung der obigen Kubischen Spline Basis

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

wobei die i-te Zeile der Modellmatrix $\mathbf{X}$ dem transponierten Prädiktorvektor

$$\mathbf{x}_i^t = \big(1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), \ldots, R(x_i, x_{q-2}^*)\big)$$

entspricht. Ist q gegeben, so können die Parameter des Modell wieder mit der Kleinsten Quadrate Methode geschätzt werden.

Zur Veranschaulichung der Kubischen Spline Basis wird das nachfolgende Beispiel betrachtet, siehe Abbildung 3.2: Für eine Rang 5 Kubische Spline Basis sind die drei Knoten $x_1^* = 1/6$, $x_2^* = 3/6$ und $x_3^* = 5/6$ gegeben. Die dafür resultierenden Basisfunktionen, $b_j(x)$, werden in den ersten fünf Bildern dargestellt. Nach Multiplikation jeder dieser Basisfunktionen mit einem Parameter, β_j, und darauffolgendem Aufsummieren ergibt dies die Funktion $f(x)$, ein Beispiel hierfür ist im unteren rechten Bild angegeben.

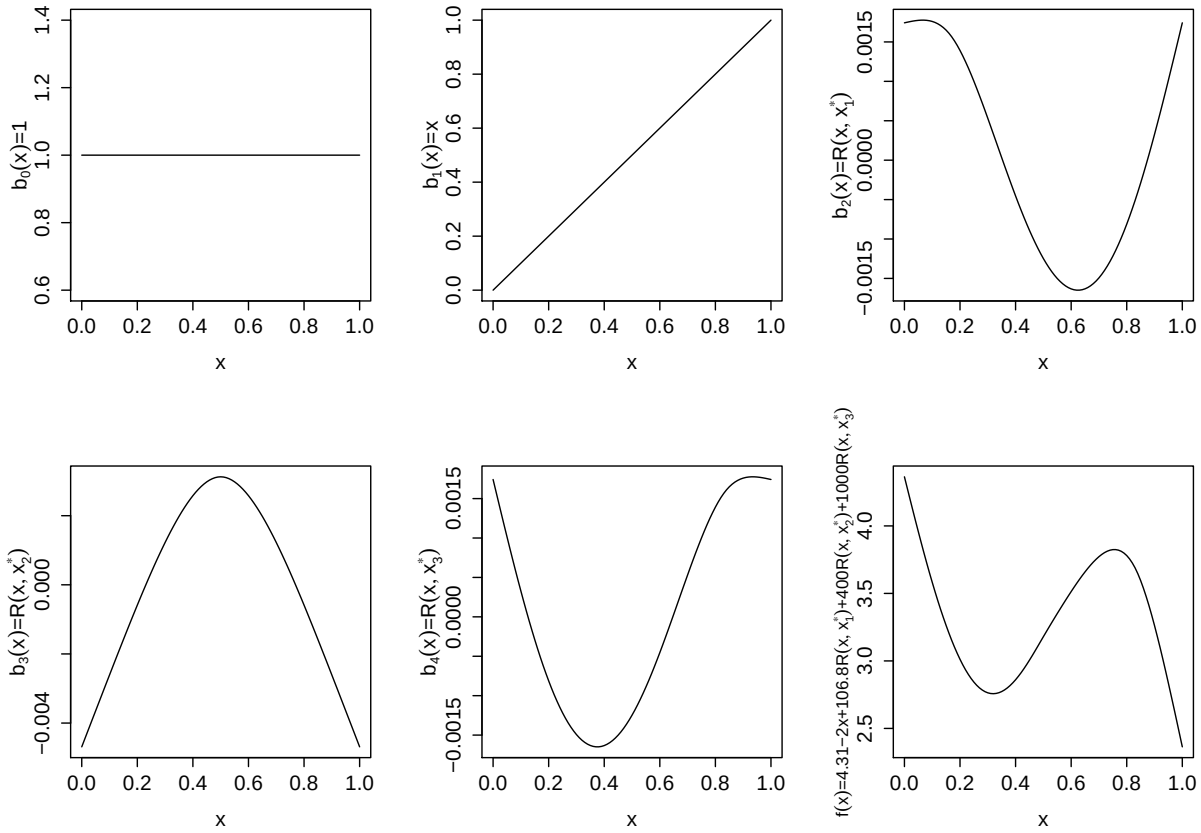

Abbildung 3.2: Die fünf Basisfunktionen einer Rang 5 Kubischen Spline Basis und eine mögliche resultierende Funktion $f(x) = 4.31 - 2x + 106.8R(x, x_1^*) + 400R(x, x_2^*) + 1000R(x, x_3^*)$ (Wood, 2006).

Anwendungsbeispiel

Um das Verhalten der zwei oben erwähnten Basen, die polynomiale Basis und die Kubische Spline Basis, bei realen Daten zu erläutern, betrachten wir die bekannten Motorradunfall-Daten von Silverman (1985). Der Datensatz, `mcycle`, kann in `R` unter Verwendung des Pakets `MASS` (siehe Venables und Ripley, 2002) aufgerufen werden und beinhaltet Werte der am Kopf gemessenen Beschleunigung nach einem simulierten Motorradunfall. Die Daten werden in Abbildung 3.3 angezeigt. Man erkennt sofort, dass kein linearer Zusammenhang zwischen Zeit und Beschleunigung vorliegt, also könnte (3.2) ein angemessenes Modell sein. Wir versuchen daher, das additive Modell zuerst unter Benutzung von polynomialen Basisfunktionen der Ordnung $q - 1$ zu schätzen.

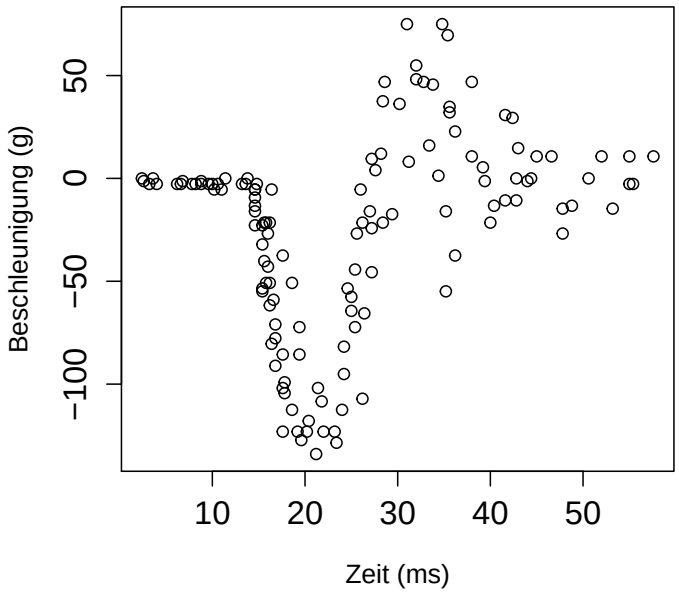

Abbildung 3.3: Die Daten von simulierten Motorradunfällen (Silverman, 1985).

```
> library(MASS)
> attach(mcycle)
> x <- times-min(times)
> x <- x/max(x)
> q <- 4
> X1 <- outer(x, 0:(q-1), "^")
> mod.poly1 <- lm(accel~X1-1)
> plot(x, accel, xlab ="skalierte Zeit", ylab="Beschleunigung (g)",
+      main="q-1=3", cex.lab=0.8, cex.main=0.8)
> lines(x, fitted(mod.poly1), col="red")
```

Der obige R-Code generiert den linken oberen Plot der Abbildung 3.4. Nachdem die x-Werte „Zeit“ skaliert wurden, damit sie in $[0, 1]$ liegen, werden die Parameter des Linearen Modells mit Designmatrix `X1` mittels der Funktion `lm` geschätzt. Da `X1` den Intercept bereits beinhaltet, muss er nicht hinzugefügt werden, deswegen `-1`. Man sieht, dass bei der Wahl von $q - 1 = 3$, keine befriedigende Schätzung des Erwartungswertes der Beschleunigung erzeugt wird. Aus diesem Grund erhöhen wir die polynomiale Ordnung, da so eine größere Flexibilität des Modells erzielt werden kann. Der rechte obere Plot der Abbilung 3.4 zeigt die Resultate der Schätzung bei einer Wahl von $q-1 = 5$, die beiden unteren Plots die bei $q - 1 = 20$ bzw. $q - 1 = 50$. Es lässt sich erkennen, dass bei einer zu hohen polynomialen Ordnung und in einem Bereich, wo wenige Beobachtungen vorhanden sind, der geschätzte Graph viel zu wackelig ist. Insgesamt scheint es so, als wäre mittels der polynomialen

Basis für diesen Datensatz keine optimale Schätzung möglich. Also wird nachfolgend die Anwendung der Kubischen Spline Basis versucht.

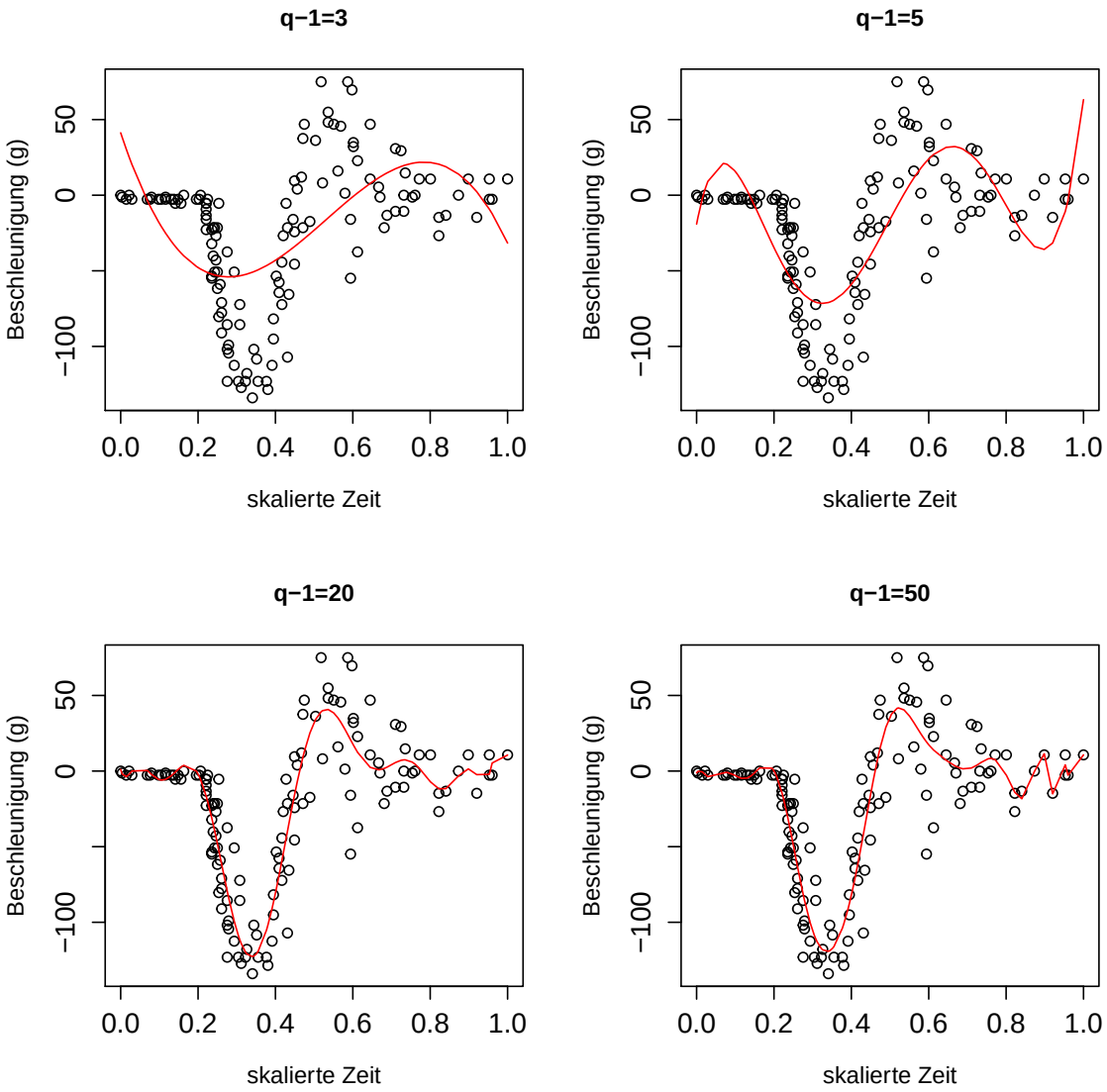

Abbildung 3.4: Der geschätzte Erwartungswert der Beschleunigung bei Anwendung polynomialer Basisfunktionen unterschiedlicher Ordnung. (Datensatz=`mcycle`)

Im ersten Schritt wird eine Funktion für $R(x, z)$ geschrieben.

```
> R <- function(x,z){
+   ((z-0.5)^2-(1/12))*((x-0.5)^2-1/12)/4-
+   ((abs(x-z)-0.5)^4-0.5*(abs(x-z)-0.5)^2+7/240)/24
+ }
```

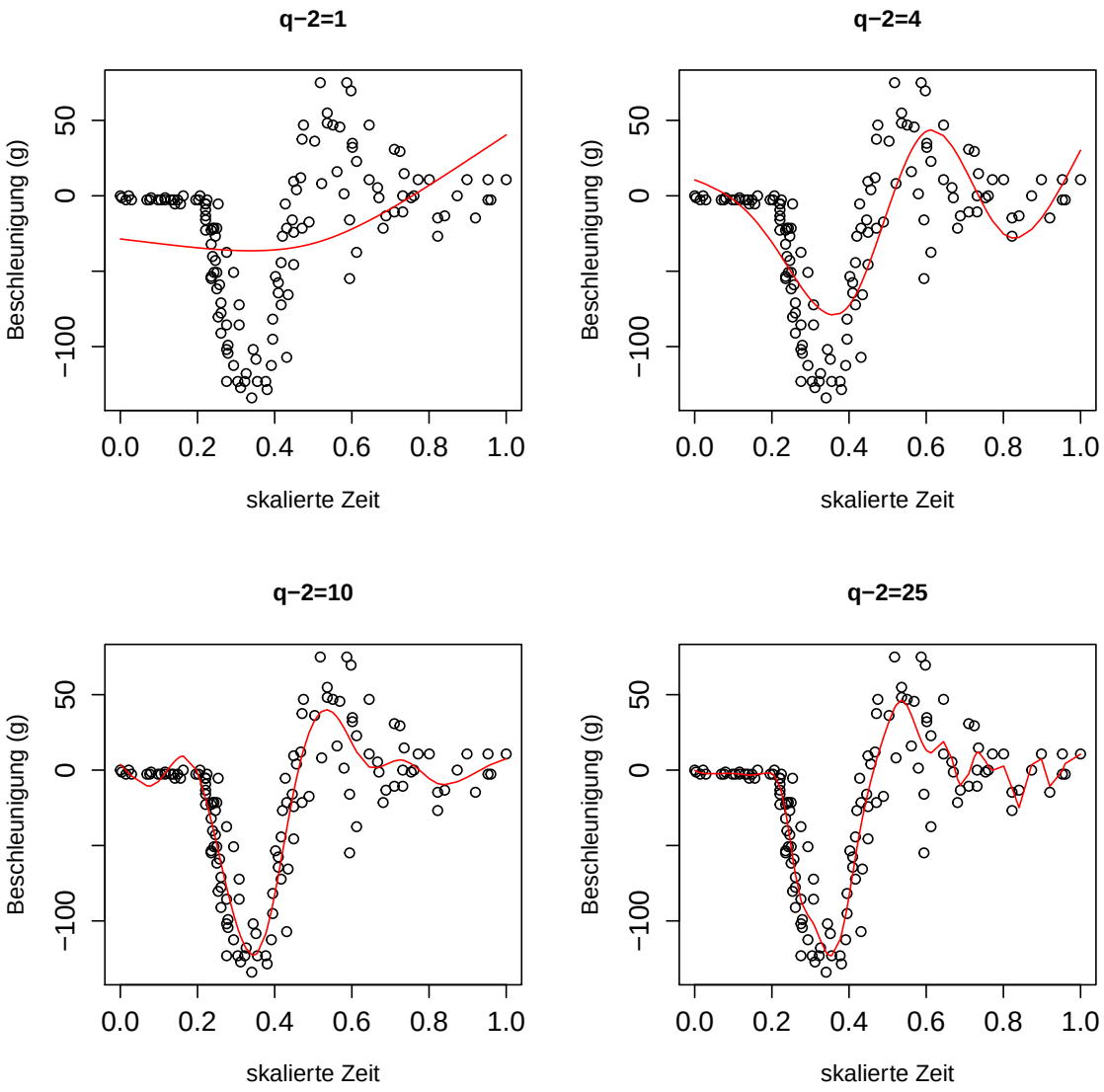

Abbildung 3.5: Der geschätzte Erwartungswert der Beschleunigung bei Gebrauch der Kubischen Spline Basis mit unterschiedlicher Knotenanzahl. (Datensatz=`mcycle`)

Unter Verwendung dieser ist es dann möglich, eine Funktion zu definieren, welche die x-Werte und eine Folge von Knoten nimmt, um die entsprechende Designmatrix zu generieren.

```
> kub.spl.X <- function(x,xm){
+   q <- length(xm)+2               # Anzahl der Parameter
+   n <- length(x)                  # Anzahl der x-Werte (Daten)
+   X <- matrix(1,n,q)              # nxq Matrix mit lauter Einser
+   X[,2] <- x                      # 2. Spalte: x-Werte
+   X[,3:q] <- outer(x, xm, FUN=R)  # Füllung der Spalten 3 bis q mit R(x,xm)
```

```
+   X
+ }
```

Jetzt benötigt man lediglich die Menge der Knoten und das Modell kann gefittet werden. Wir versuchen es mal mit nur einem Knoten, $x_1^* = 1/2$.

```
> xm <- 1/2
> X1.kub <- kub.spl.X(x,xm)
> mod.kub1 <- lm(accel~X1.kub-1)
> plot(x, accel, xlab ="skalierte Zeit", ylab="Beschleunigung (g)",
+      main="q-2=1", cex.lab=0.8, cex.main=0.8)
> lines(x, fitted(mod.kub1))
```

Bei der Ausführung des obigen R-Codes erhält man den linken oberen Plot der Abbildung 3.5. Er zeigt die geschätzten Erwartungswerte der Beschleunigung. Die restlichen Grafiken dieser Abbildung resultieren von einer Erhöhung der Anzahl der Knoten, $q-2$, auf $4, 10$ bzw. 25. Man kann beobachten, dass die Knotenzahl die gefitteten Werte stark beeinflusst. Während bei einer geringen Anzahl von Knoten eine unbefriedigende Anpassung erfolgt (linker oberer Plot), so führt eine hohe Knotenzahl zu einer wackeligen Anpassung (rechter unterer Plot). Wir sind also daran interessiert, den Grad der Glätte, welcher hier durch die Anzahl der Knoten (bei polynomialer Basis durch die polynomiale Ordnung $q-1$) kontrolliert wird, so zu wählen, dass die Anpassung optimal wird. Dies führt uns zur zweiten Frage: Wie glatt soll f sein?

3.1.2 Penalized Regression Splines

Eine Möglichkeit (siehe z.B. Eilers und Marx, 1996; Gu, 2002; Wood, 2006; Fahrmeir et al., 2009) den idealen Grad der Glätte zu bestimmen, ist die Dimension der Basis, q (bei der Kubischen Spline Basis: Anzahl der Knoten + 2), fix zu halten, die gewählte Größe soll die wahrscheinlich notwendige Dimension ein wenig übersteigen, jedoch einen Strafterm für die Kontrolle der Glätte hinzuzufügen. D.h., im Gegensatz zu Abschnitt 2.1 wird nun nicht

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

bezüglich $\boldsymbol{\beta}$ minimiert, sondern

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \int_0^1 \left(f''(x)\right)^2 dx, \tag{3.5}$$

wobei der zweite Term Modelle bestraft, die zu wackelig sind, und λ der Glättungsparameter (smoothing parameter) ist. Während bei $\lambda \to \infty$ die Schätzung von f zu einer Geraden wird, so folgt bei $\lambda = 0$ ein Modell ohne Strafterm. Es sei noch erwähnt, dass anstelle von $\int_0^1 \left(f''(x)\right)^2 dx$ auch andere Strafterme gewählt werden können, wie z.B. die Differenzen-Strafterme

$$\sum_{j=1}^{q-1} (\beta_j - \beta_{j-1})^2 \qquad \text{oder} \qquad \sum_{j=1}^{q-2} (\beta_{j+1} - 2\beta_j + \beta_{j-1})^2 .$$

Detailliertere Informationen können in Eilers und Marx (1996) nachgelesen werden. In weiterer Folge wird jedoch der Strafterm in (3.5) verwendet.

Da f linear in den Parametern, β_j, ist, kann der Strafterm auch als

$$\int_0^1 \left(f''(x)\right)^2 dx = \boldsymbol{\beta}^t \mathbf{S} \boldsymbol{\beta}$$

geschrieben werden (Wood, 2006), wobei $\mathbf{S}$ eine Matrix mit bekannten Koeffizienten ist. Werden die Kubischen Spline Basen als Basisfunktionen gewählt, so stellt sich heraus, dass $S_{i+2,j+2} = R(x_i^*, x_j^*)$ für $i, j = 1, \ldots, q-2$ und die Einträge der ersten zwei Zeilen und Spalten 0 sind (Gu, 2002). Also bedeutet die Schätzung des Modells bei Anwendung der Penalisierten Kleinsten Quadrate Methode (penalized least squares method) die Minimierung von

$$\mathbf{S}_p := \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^t \mathbf{S} \boldsymbol{\beta}. \tag{3.6}$$

Die Bestimmung des Glättungsgrades ist somit nicht ein Problem der Schätzung der Dimension der Basis q, sondern die des Glättungsparameters λ. Doch bevor wir uns mit der Wahl von λ beschäftigen, konzentrieren wir uns auf die Schätzung von $\boldsymbol{\beta}$ bei gegebenem λ.

Um (3.6) bezüglich $\boldsymbol{\beta}$ zu minimieren wird zuerst $\mathbf{S}_p$ umgeformt:

$$\begin{aligned}
\mathbf{S}_p &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^t \mathbf{S} \boldsymbol{\beta} \\
&= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^t \mathbf{S} \boldsymbol{\beta} \\
&= \mathbf{y}\mathbf{y}^t - 2\boldsymbol{\beta}^t \mathbf{X}^t \mathbf{y} + \boldsymbol{\beta}^t \left(\mathbf{X}^t \mathbf{X} + \lambda \mathbf{S}\right) \boldsymbol{\beta}.
\end{aligned}$$

Nach der Ableitung von $\mathbf{S}_p$ nach $\boldsymbol{\beta}$

$$\frac{\partial \mathbf{S}_p}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^t \mathbf{y} + 2\mathbf{X}^t \mathbf{X} \boldsymbol{\beta} + 2\lambda \mathbf{S} \boldsymbol{\beta}$$

und anschließendem Nullsetzen erhalten wir den gesuchten Schätzer $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t \mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^t \mathbf{y}.$$

Für die gefitteten Werte ergibt sich also

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

wobei die sogenannte Hatmatrix $\mathbf{A}$ die Definition

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^t \mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^t$$

besitzt.

Jedoch weist $\hat{\boldsymbol{\beta}}$ bei direktem Verwenden eine suboptimale numerische Stabilität auf. Deswegen wird für praktische Berechnungen der folgende Term herangezogen:

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

wobei für $\mathbf{B}$ gilt, dass $\mathbf{B}^t\mathbf{B} = \mathbf{S}$. Die linke Seite des obigen Ausdrucks kann als erweitertes Kleinste Quadrate Problem gesehen werden, bei welchem die Responsevariable mit q Nullen und die Designmatrix mit $\sqrt{\lambda}\mathbf{B}$ ergänzt wurde. Was noch fehlt, ist die Berechnung von $\mathbf{B}$, diese kann z.B. mittels der Spektralzerlegung oder der Choleskyzerlegung bestimmt werden (siehe Wood, 2006). Nach erfolgter Ermittlung kann das obige erweiterte Kleinste Quadrate Problem sehr einfach mit den gängigen Methoden der Linearen Regressionsanalyse gelöst werden.

Um nun Penalisierte Regression Splines in der Anwendung zu sehen, ziehen wir wieder die Motorradunfall-Daten (`mcylce`) heran und schreiben als Erstes eine Funktion, welche die Matrix $\mathbf{S}$ berechnet.

```
> pen.S <- function(xm){
+   q <- length(xi)+2
+   S <- matrix(0,q,q)
+   S[3:q,3:q] <- outer(xm, xm, FUN=R)
+   S
+ }
```

Unter Verwendung der Spektralzerlegung erzeugen wir im nächsten Schritt $\mathbf{B} = \sqrt{\mathbf{S}}$ (siehe A.8, Wood (2006)).

```
> sqrt.S <- function(S){
+   r <- eigen(S,symmetric=TRUE)
+   B <- r$vectors%*%diag(r$values^0.5)%*%t(r$vectors)
+   B
+ }
```

Zum Schluss muss noch die erweiterte Responsevariable und die erweiterte Designmatrix generiert werden.

```
> pen.spl.X <- function(x,xm,lambda){
+   X.pen <- rbind(kub.spl.X(x,xm), sqrt.S(pen.S(xm))*sqrt(lambda))
+   X.pen
+ }
> pen.spl.y <- function(y,xm){
+   q <- length(xm)+2
+   n <- length(y)
+   y.pen <- y
+   y.pen[(n+1):(n+q)] <- 0
+   y.pen
+ }
```

Nun sind alle Funktionen definiert um das Modell fitten zu können. Alles, was noch benötigt wird, sind die Eingabewerte, nämlich die Anzahl der Knoten $q-2$ bzw. die Dimension der Basis q, die Position der Knoten x_i^* und der Wert des Glättungsparameters λ. Bei der Wahl der Basisdimension ist zu beachten, dass dieser ein wenig größer als die vermutete sein soll, also wählen wir $q = 22$ (siehe Abbildung 3.5). Demzufolge ist die Knotenzahl 20,

wobei diese gleichmäßig auf $[0, 1]$ positioniert werden. Da der Glättungsparameter für die Flexibilität des Modells zuständig ist, werden in weiterer Folge einige Werte für λ getestet. Der nachstehende R-Code schätzt die unbekannten Parameter β_j für $\lambda = 0.1$ und erzeugt die linke obere Grafik der Abbildung 3.6.

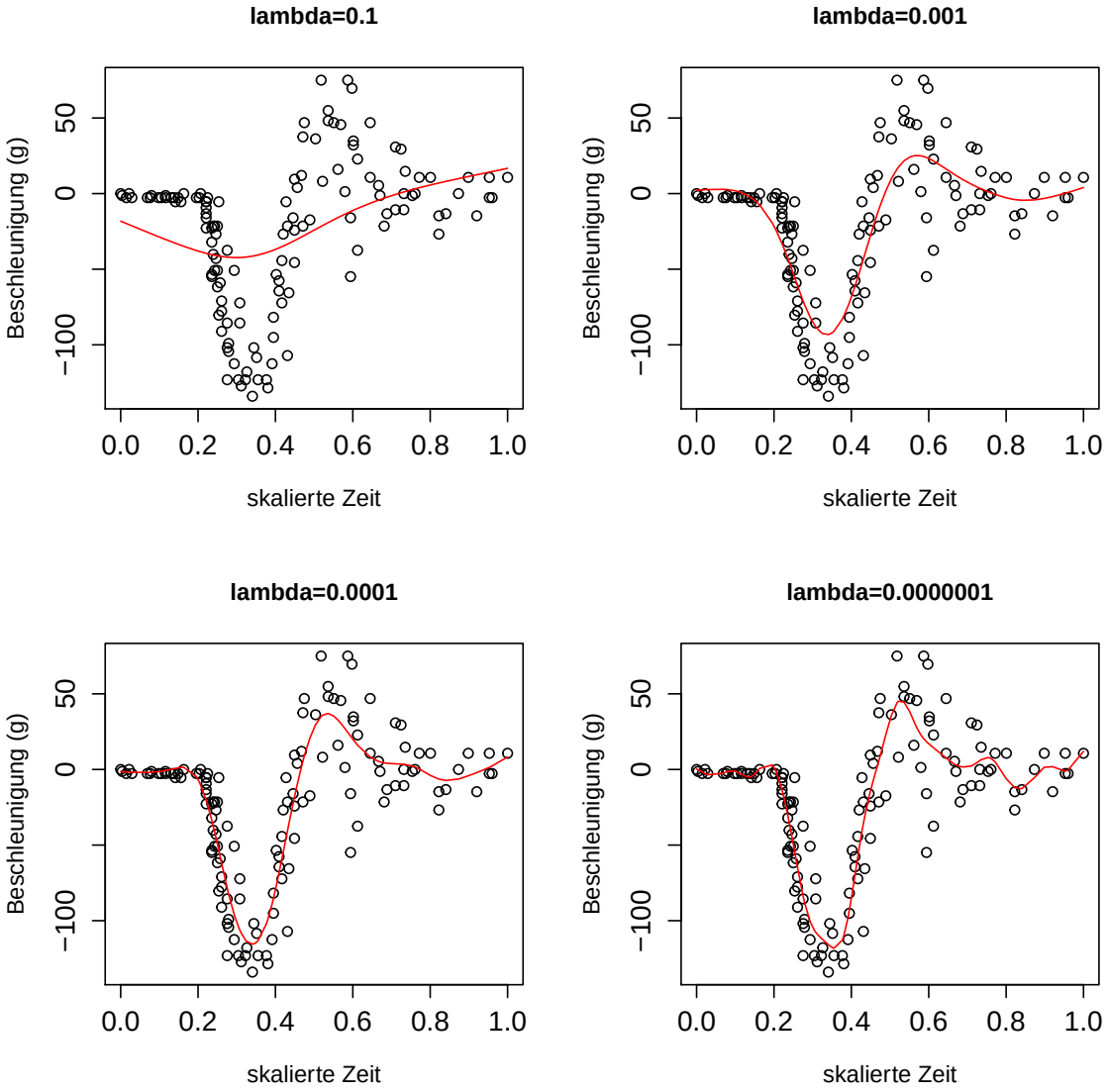

Abbildung 3.6: Der geschätzte Erwartungswert der Beschleunigung bei Anwendung Penalisierter Regression Splines mit unterschiedlichen Glättungsparameter λ. (Datensatz=`mcycle`)

```
> xm <- 1:20/21
> X1.pen <- pen.spl.X(x,xm,0.1)
> y.pen <- pen.spl.y(accel,xm)
> mod.pen1 <- lm(y.pen~X1.pen-1)
> plot(x, accel, xlab ="skalierte Zeit", ylab="Beschleunigung (g)",
```

```
+      main="lambda=0.1", cex.lab=0.8, cex.main=0.8)
> X.kub <- kub.spl.X(x,xm)
> lines(x, X.kub%*%coef(mod.pen1))
```

Die restlichen Plots resultieren aus einer verminderten Wahl von λ, nämlich 0.001, 0.0001 und 10^{-7}. Man sieht sofort, dass bei einem kleineren Glättungsparameter die Anpassung wackeliger ist und umgekehrt, je größer λ, desto ungenauer die Schätzung. Also, wie groß soll λ nun gewählt werden?

3.1.3 Wahl des Glättungsparameters

Eine Möglichkeit, um den Glättungsparamter zu wählen, wäre

$$M = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - f(x_i)\right)^2$$

zu minimieren, das bedeutet, λ so zu wählen, dass der quadrierte Abstand von $\hat{f}$ zu f minimal wird. Dieser Ausdruck kann aber nicht berechnet werden, da f unbekannt ist; man versucht jedoch, daraus einen sinnvollen Schätzer abzuleiten. Dazu betrachtet man die sogenannte Kreuzvalidierung. Bei dieser wird jeweils eine Beobachtung aus den Daten gestrichen, dann die Schätzung für f mit den verbleibenden Daten durchgeführt, sei dies $\hat{f}^{[-i]}(x)$, und schließlich die quadrierte Differenz zwischen der fehlenden Beobachtung y_i und deren Vorhersagewert $\hat{f}^{[-i]}(x_i)$ ermittelt. Also definiert sich das Kreuzvalidierungskriterium (CV-Kriterium) durch

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}^{[-i]}(x_i) - y_i\right)^2.$$

Doch wie kann ein Zusammenhang mit M hergestellt und somit das CV-Kriterium theoretisch gerechtfertigt werden? Hierfür ersetzen wir $y_i = f(x_i) + \epsilon_i$ in CV und wegen $\mathbb{E}(\epsilon_i) = 0$ und der Unabhängigkeit von ϵ_i und $\hat{f}^{[-i]}(x_i)$ folgt für den Erwartungswert des CV-Kriteriums

$$\begin{aligned}
\mathbb{E}(\mathrm{CV}) &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(\hat{f}^{[-i]}(x_i) - f(x_i) - \epsilon_i\right)^2\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n}\left(\hat{f}^{[-i]}(x_i) - f(x_i)\right)^2 - 2\left(\hat{f}^{[-i]}(x_i) - f(x_i)\right)\epsilon_i + \epsilon_i^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}\left(\left(\hat{f}^{[-i]}(x_i) - f(x_i)\right)^2\right) - 2\,\mathbb{E}\left(\hat{f}^{[-i]}(x_i) - f(x_i)\right)\mathbb{E}\left(\epsilon_i\right) + \mathbb{E}\left(\epsilon_i^2\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left(\hat{f}^{[-i]}(x_i) - f(x_i)\right)^2\right) + \sigma^2.
\end{aligned}$$

Und da $\hat{f}^{[-i]} \approx \hat{f}$ gilt, ergibt sich $\mathbb{E}(\text{CV}) \approx \mathbb{E}(M) + \sigma^2$ und somit ein akzeptabler Konnex zwischen CV und M. Also ist die Minimierung des CV-Kriteriums eine vertretbare Vorgangsweise um λ zu wählen.

Zur Bestimmung des CV-Kriteriums sind beim ersten Anblick n separate Schätzungen des Regressionsmodells notwendig. Man kann jedoch zeigen, dass

$$\text{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{f}(x_i) - y_i}{1 - A_{ii}}\right)^2,$$

wobei $\mathbf{A}$ die dazugehörige Hatmatrix ist. D.h., die Berechnung des CV-Kriteriums ist auch ohne die Schätzungen von $\hat{f}^{[-i]}(x_i)$ möglich und folglich numerisch effizienter.

In der Praxis werden die Gewichte, $1 - A_{ii}$, häufig durch ihren Mittelwert, $sp(\mathbf{I} - \mathbf{A})/n$, ersetzt, sodass das generalisierte Kreuzvalidierungskriterium (GCV-Kriterium) resultiert:

$$\text{GCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{f}(x_i) - y_i}{sp(\mathbf{I} - \mathbf{A})/n}\right)^2 = \frac{n\sum_{i=1}^{n}\left(\hat{f}(x_i) - y_i\right)^2}{\left(sp(\mathbf{I} - \mathbf{A})\right)^2}.$$

Dabei bezeichnet $\text{sp}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii}$ die Spur der Matrix $\mathbf{A}$. Die Verwendung des GCV-Kriteriums birgt einige Vorteile, denn neben der leichteren Berechenbarkeit im Vergleich zum CV-Kriterium weist es einen theoretischen Vorteil in Bezug auf die Invarianz auf (vgl. Golub, Heath und Wahba, 1979).

Zum Abschluss dieses Abschnitts kommen wir noch einmal zu dem Motorradunfall-Beispiel (`mcycle`) zurück und versuchen den optimalen Glättungsparameter mit dem generalisierten Kreuzvalidierungskriterium zu ermitteln. Dafür konstruieren wir zuerst eine Schleife, welche für steigende λ-Werte den dazugehörigen GCV-Wert kalkuliert. Anschließend werden die berechneten GCV-Werte graphisch dargestellt (siehe linker Plot Abbildung 3.7).

```
> xi <- 1:20/21
> y.pen <- pen.spl.y(accel,xi)
> lambda <- 10^(-9)
> n <- length(x)
> GCV <- 0

> for(i in 1:65){
+   X.pen <- pen.spl.X(x,xi,lambda)
+   mod <- lm(y.pen~X.pen-1)
+   spA <- sum(influence(mod)$hat[1:n])
+   sse <- sum((accel-fitted(mod)[1:n])^2)
+   GCV[i] <- (n*sse)/(n-spA)^2
+   lambda <- lambda*1.5
+ }
> plot(1:65, GCV, type="l", xlab="i", ylab="GCV", main="GCV-Werte",
+      cex.lab=0.8, cex.main=0.8)
```

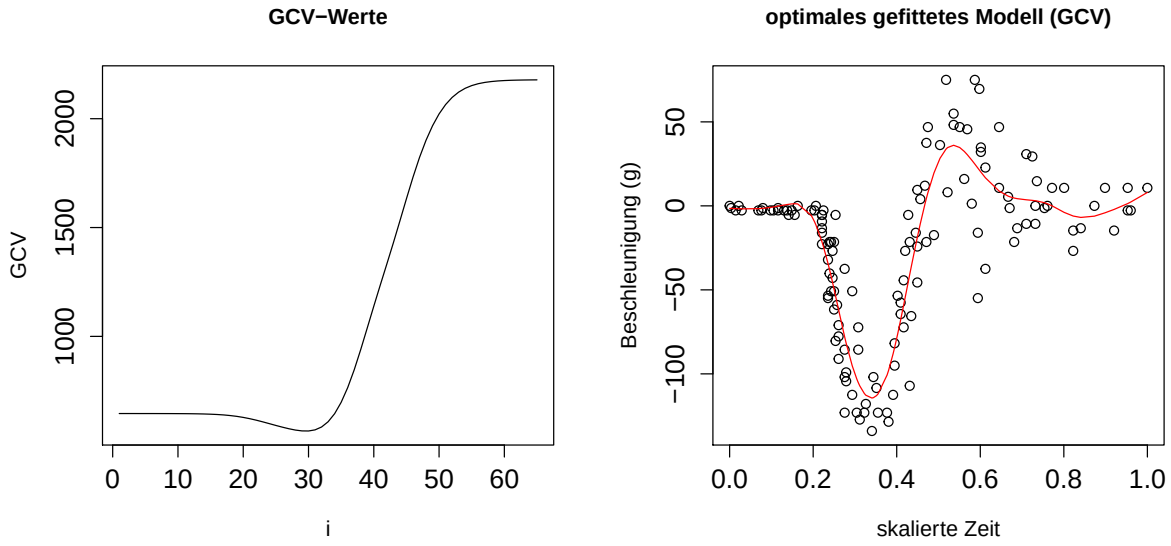

Abbildung 3.7: Links die GCV-Funktion. Rechts der geschätzte Erwartungswert der Beschleunigung bei Gebrauch des optimalen Glättungsparameters (laut GCV-Kriterium). (Datensatz=`mcycle`)

Nun muss die Stelle des `GCV`-Vektors gefunden werden, an welcher dieser den kleinsten Wert aufweist, um im Anschluss den optimalen Glättungsparameter bestimmen zu können.

```
> imin <- 0
> for(i in 1:65){
+   if(GCV[i]==min(GCV)){
+     imin <- i
+   }
+ }
> imin
[1] 30
> lambdaopt <- 10^(-9)*1.5^(imin-1)
> lambdaopt
[1] 0.000127834
```

Der niedrigste Wert wird also bei `GCV[30]` erreicht und somit ist der optimale Glättungsparameter $\hat{\lambda} = 10^{-9} \cdot 1.5^{29} \approx 1.28 \cdot 10^{-4}$. Was noch fehlt, ist die Schätzung des Modells mit dem optimalen λ-Wert und deren graphische Darstellung (siehe rechter Plot Abbildung 3.7).

```
> X.pen.opt <- pen.spl.X(x,xi,lambdaopt)
> mod.opt <- lm(y.pen~X.pen.opt-1)
> plot(x, accel, xlab ="skalierte Zeit", ylab="Beschleunigung (g)",
+      main="optimales gefittetes Modell (GCV)", cex.lab=0.8, cex.main=0.8)
> X.kub <- kub.spl.X(x,xi)
> lines(x, X.kub%*%coef(mod.opt), col="red")
```

3.2 Generalisierte Additive Modelle

Nach der Einführung der Additiven Modelle fokussieren wir uns nun auf deren Erweiterung, die Klasse der Generalisierten Additiven Modelle (GAM). Dieser Abschnitt stützt sich wieder auf Wood (2006). Für weitere Erklärungen bzw. alternative Herangehensweisen siehe z.B. Hastie und Tibshirani (1986, 1990), Marx und Eilers (1998) oder Ruppert et al. (2003).

Wie bereits am Anfang des Abschnitts 3.1 erwähnt, besitzt ein GAM beispielsweise die Struktur

$$g(\mu_i) = \mathbf{x}_i^{*\,t}\boldsymbol{\beta} + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}, x_{i4}) + \ldots \tag{3.7}$$

mit

$$\mu_i = \mathbb{E}(y_i), \qquad y_i \overset{ind}{\sim} \text{Exponentialfamilie}(\theta_i).$$

Hierbei ist y_i die Responsevariable, $\mathbf{x}_i^*$ stellt den Prädiktorvektor für die parametrischen Modellkomponenten dar, $\boldsymbol{\beta}$ bezeichnet den Parametervektor und $f_k(\cdot)$ ist eine glatte Funktion der Prädiktorvariablen.

Unser Ziel ist es nun ein solches Modell zu schätzen, wobei im Folgenden angenommen wird, dass jede glatte Funktion $f_k(\cdot)$ univariat ist. Beinhaltet das betrachtete Modell mehrere univariate Glättungsterme, so ist das Modell nicht identifizierbar; es liegt also ein Identifikationsproblem vor. Dies lässt sich am folgenden Beispiel (siehe Fahrmeir et al., 2009) erklären: Wird einer Funktion $f_1(x_1)$ die Konstante $c \neq 0$ hinzuaddiert und zeitgleich c einer zweiten Funktion $f_2(x_2)$ abgezogen, so vollzieht sich die Summe

$$f_1(x_1) + f_2(x_2) = f_1(x_1) + c + f_2(x_2) - c = \tilde{f}_1(x_1) + \tilde{f}_2(x_2)$$

keiner Veränderung, wobei $\tilde{f}_1(x_1) = f_1(x_1) + c$ und $\tilde{f}_2(x_2) = f_2(x_2) - c$. Demnach besitzen die Funktionen $f_1(x_1)$ und $\tilde{f}_1(x_1)$ bzw. $f_2(x_2)$ und $\tilde{f}_2(x_2)$ die gleiche Form, sind aber nicht identifizierbar. Um dieses Problem zu beheben wird oftmals eine Zentrierungsnebenbedingung eingeführt, bei welcher die Funktionen um Null zentriert werden, d.h.

$$\sum_{i=1}^{n} f_k(x_{ik}) = 0$$

für $k = 1, \ldots, d$. Für die Glättungsterme f_k müssen also die Basisfunktionen und die Parameter derart gewählt werden, sodass die Zentrierungsnebenbedingung erfüllt ist (für genauere Informationen siehe Wood, 2006). Seien die passenden Vektoren der Basisfunktionen $\mathbf{b}_k = (b_{k1}(\cdot), \ldots, b_{kq_k}(\cdot))^t$, $k = 1, \ldots, d$, so ergibt sich für jedes f_k

$$f_k(\cdot) = \sum_{j=1}^{q_k} b_{kj}(\cdot)\gamma_{kj} = \mathbf{b}_k^t\boldsymbol{\gamma}_k,$$

wobei $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kq_k})^t$ der unbekannte Parametervektor ist, welcher geschätzt werden muss. Wichtig ist es hier zu erwähnen, dass bis zu dieser Stelle der Arbeit die Laufvariable j immer bei Null gestartet ist, jedoch in diesem Abschnitt eine absichtliche Verschiebung

stattfindet, um ein leichteres Verständnis der kommenden Definitionen zu ermöglichen. Betrachten wir nun den nichtparametrischen Teil von (3.7), so erhalten wir für jedes $i = 1, \dots, n$

$$f_1(x_{i1}) + \cdots + f_d(x_{id}) = \mathbf{b}_1^t \boldsymbol{\gamma}_1 + \cdots + \mathbf{b}_d^t \boldsymbol{\gamma}_d.$$

Völlig äquivalent dazu kann dies in Vektorschreibweise als

$$\mathbf{f}_1 + \cdots + \mathbf{f}_d = \mathbf{B}_1 \boldsymbol{\gamma}_1 + \cdots + \mathbf{B}_d \boldsymbol{\gamma}_d$$

dargestellt werden, mit $\mathbf{f}_k = (f_k(x_{1k}), \dots, f_k(x_{nk}))^t$ für jedes $k = 1, \dots, d$ und der k-ten Designmatrix

$$\mathbf{B}_k = \begin{pmatrix} b_{k1}(x_{1k}) & \dots & b_{kq_k}(x_{1k}) \\ \vdots & & \vdots \\ b_{k1}(x_{nk}) & \dots & b_{kq_k}(x_{nk}) \end{pmatrix}.$$

Insgesamt kann also ein GAM der Form (3.7) in

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\gamma} \tag{3.8}$$

umgeschrieben werden, mit Designmatrix $\mathbf{X} = (\mathbf{X}^* | \mathbf{B}_1 | \dots | \mathbf{B}_d)$ und Parametervektor $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^t, \boldsymbol{\gamma}_1^t, \dots, \boldsymbol{\gamma}_d^t\right)^t$. Dabei entspricht $\mathbf{X}^*$ der Designmatrix für die parametrischen Modellkomponenten. Die obige Darstellung ist eindeutig ein GLM und die unbekannten Parameter können mittels der Maximum-Likelihood Methode geschätzt werden. Jedoch kann es sehr leicht zu einer Überanpassung kommen, falls die q_k zu groß gewählt werden. Deswegen wird für GAMs üblicherweise die penalisierte Maximum-Likelihood Methode angewandt. Die Strafterme haben die Aufgabe, die zu wackeligen Schätzungen der f_k-Terme zu unterdrücken, und besitzen z.B. die Form $\boldsymbol{\gamma}^t \mathbf{S}_k \boldsymbol{\gamma}$, wobei $\mathbf{S}_k$ eine Matrix mit bekannten Koeffizienten ist. An dieser Stelle muss noch erwähnt werden, dass die Strafterme bereits so gewählt sind, dass diese der Zentrierungsnebenbedingung gerecht werden (unter Verwendung einer Reparametrisierung ist dies jederzeit realisierbar, siehe Wood, 2006).

Nun ist es möglich, die penalisierte Log-Likelihoodfunktion für ein GAM zu definieren:

$$l_p(\boldsymbol{\gamma}|\mathbf{y}) = l(\boldsymbol{\gamma}|\mathbf{y}) - \frac{1}{2} \sum_{k=1}^{d} \lambda_k \boldsymbol{\gamma}^t \mathbf{S}_k \boldsymbol{\gamma},$$

wobei $l(\boldsymbol{\gamma}|\mathbf{y})$ die Log-Likelihoodfunktion von (3.8) ist und λ_k die Glättungsparameter symbolisiert. Sind also die Werte von λ_k gegeben, so kann l_p maximiert werden um den gesuchten Schätzer $\hat{\boldsymbol{\gamma}}$ zu erhalten. Doch bevor wir uns mit der Lösung der penalisierten Log-Likelihoodfunktion beschäftigen, richten wir unsere Konzentration auf die Wahl einer möglichen Basis.

3.2.1 Basen

Im vorigen Kapitel wurden der Einfachheit halber nur zwei Basen eingeführt, allerdings existieren in der Praxis viele mehr. Die meisten der verwendeten glatten Funktionen beruhen auf einer Spline Basis, da Splines einige vorteilhafte theoretische Eigenschaften besitzen. Zur Wiederholung sei ihre Definition nochmal angeführt: Eine Funktion $f : [a, b] \to \mathbb{R}$

nennt man Polynomialen Spline vom Grad $l \geq 0$ zu den Knoten $a = x_1^* < \cdots < x_m^* = b$, falls sie $(l-1)$-mal stetig differenzierbar ist und sich stückweise aus Polynomen vom Grad l, eine für jedes Intervall $\left[x_j^*, x_{j+1}^*\right]$, zusammensetzt, welche sich an den Knoten verbinden. Wird zusätzlich angenommen, dass $f''(x_1^*) = f''(x_m^*) = 0$ gilt, dann nennt man die Funktion $f(x)$ natürlicher polynomialer Spline.

Die einfachste Möglichkeit für die Wahl einer Basis ist der Lineare Spline. Nach der Analyse dieser werden wir uns der Untersuchung der Kubischen Splines widmen und sehen, dass jene eine Minimalitätsbedingung erfüllen, welche sie im Vergleich zu anderen Splines besonders interessant macht. Zu guter Letzt werden wir uns mit B-Splines bzw. P-Splines beschäftigen. Neben all diesen Basen liegen noch viele weitere vor, wie z.B. der Thin Plate Spline oder die Tensor Produkt Basis für glatte Interaktionseffekte, welche jedoch in dieser Arbeit nicht behandelt werden. Für weitere Informationen empfiehlt es sich beispielsweise Duchon (1977), de Boor (1978) oder Wood (2006) zu betrachten.

Lineare Spline Basis
Angenommen, es sind m Knoten, $x_1^*, \ldots, x_m^*$, gegeben, dann besitzt die Lineare Spline Funktion die folgende Definition:

$$f(x) = \sum_{j=1}^{m} b_j(x)\gamma_j, \tag{3.9}$$

wobei die Basisfunktionen Dreiecksfunktionen sind. Dabei hat die j-te Dreiecksfunktion den Wert Eins beim Knoten x_j^* und fällt linear bis Null beim Knoten x_{j-1}^* bzw. x_{j+1}^*. Anderweitig ist die Funktion Null. Also lässt sich die j-te Basisfunktion folgenderweise schreiben:

$$b_j(x) = \begin{cases} 0, & \text{falls } x \leq x_{j-1}^* \text{ und } x \geq x_{j+1}^*, \\ \frac{x - x_j^* + h_{j-1}}{h_{j-1}}, & \text{falls } x_{j-1}^* < x \leq x_j^*, \\ \frac{x_j^* - x + h_j}{h_j}, & \text{falls } x_j^* < x < x_{j+1}^*, \end{cases}$$

mit $h_j = x_{j+1}^* - x_j^*$.

Beispiel: Für die bessere Nachvollziehbarkeit der Linearen Splines wird nun ein Beispiel behandelt. Es seien 10 Knoten gegeben, $x_1^*, \ldots, x_{10}^*$, welche gleichmäßig auf $[0, 1]$ verteilt sind. Im ersten Schritt werden die 10 Basisfunktionen erzeugt (siehe Abbildung 3.8).

```
> # erzeuge j-te Basisfunktion für Knoten xm
> bj <- function(x,xm,j){
+   ym <- xm*0
+   ym[j] <- 1
+   approx(xm,ym,x)$y
+ }

> # plotte die Basisfunktionen für 10 Knoten
> m <- 10
```

```
> xm <- c(0, 1:8/9, 1)
> plot(xm, xm*0, ylim=c(0,2), xlab="x", ylab="", cex.lab=0.8)
> title(ylab=expression(paste(b[j](x))), mgp=c(2.0,1,0), cex.lab=0.8)
> for(j in 1:m){
+  lines(seq(0,1,length=10000), bj(seq(0,1,length=10000),xm,j), col=j, lty=2)
+ }
```

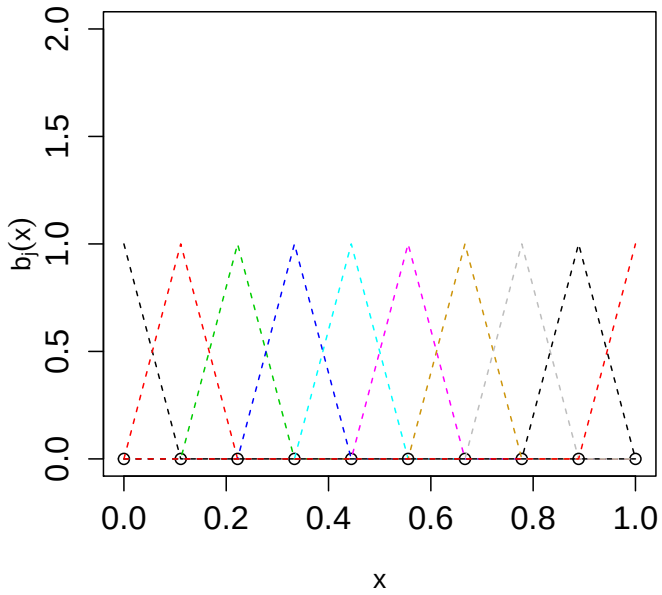

Abbildung 3.8: Die Basisfunktionen des Linearen Splines.

Nun ist es möglich, eine Lineare Spline Funktion zu erstellen. Dafür wird jede Dreiecksfunktion, $b_j(x)$, mit ihren Koeffizienten, γ_j, multipliziert und anschließend werden die Produkte aufsummiert. An den Knoten, x_j^*, besitzt also die Lineare Spline Funktion genau den Wert des j-ten Koeffizienten, also $f(x_j^*) = \gamma_j$. Zur Illustrierung wird eine mögliche Lineare Spline Funktion in Abbildung 3.9 angezeigt.

```
> # erzeuge und plotte eine mögliche lineare Spline Funktion
> f <- function(x,coef,xm,m){
+   temp <- 0
+   for(j in 1:m){
+     temp <- temp+(coef[j]*bj(x,xm,j))
+   }
+   temp
+ }
> coef <- c(1, 1.5, 1.4, 1.7, 2.1, 1.9, 2.3, 2.6, 2.8, 2.4)
```

```
> plot(xm, xm*0, ylim=c(0,3), xlab="x", ylab="f(x)", cex.lab=0.8)
> for(j in 1:m){
+   lines(seq(0,1,length=10000), coef[j]*bj(seq(0,1,length=10000),xm,j),
+         col=j, lty=2)
+ }
> lines(seq(0,1,length=10000), f(seq(0,1,length=10000),coef,xm,m), lwd=2)
```

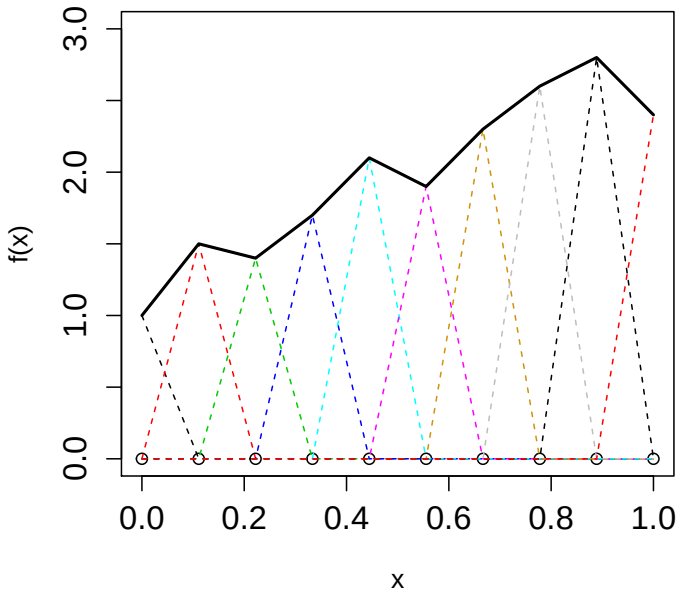

Abbildung 3.9: Eine mögliche Lineare Spline Funktion.

Kubische Spline Basis

Die Lineare Spline Basis ist nicht schlecht, jedoch können bessere und allgemeinere Basen für eine Vielzahl von Aufgabenstellungen gefunden werden. Eine häufig verwendete ist die Kubische Spline Basis, da diese den folgenden Vorteil aufweist.

Satz 3.1. Von allen Funktionen f, die auf $[a, b]$ stetig sind und absolut stetige erste Ableitungen besitzen, ist der Kubische Spline $g(x)$ die Funktion, welche

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b f''(x)^2 dx$$

minimiert, falls $\lambda \geq 0$ fix ist.

Beweis: siehe Reinsch (1967)

Nun möchten wir eine von Abschnitt 2.4 abweichende Definition der Kubischen Spline Funktion einführen, welche Vorzüge im Bereich der Interpretation der Parameter besitzt. Seien m Knoten, $x_1^*, \ldots, x_m^*$, gegeben, dann kann die Kubische Spline Funktion $f(x)$ für $j = 1, \ldots, m-1$ durch

$$f(x) = a_j^-(x)\gamma_j + a_j^+(x)\gamma_{j+1} + c_j^-(x)\delta_j + c_j^+(x)\delta_{j+1}, \qquad \text{falls } x_j^* \leq x \leq x_{j+1}^*, \tag{3.10}$$

definiert werden, wobei $\gamma_j = f(x_j^*)$ bzw. $\delta_j = f''(x_j^*)$ gilt und die Basisfunktionen die folgende Form besitzen:

$$\begin{aligned} a_j^-(x) &= \left(x_{j+1}^* - x\right)/h_j & c_j^-(x) &= \left(\left(x_{j+1}^* - x\right)^3/h_j - h_j\left(x_{j+1}^* - x\right)\right)/6 \\ a_j^+(x) &= \left(x - x_j^*\right)/h_j & c_j^+(x) &= \left(\left(x - x_j^*\right)^3/h_j - h_j\left(x - x_j^*\right)\right)/6, \end{aligned}$$

mit $h_j = x_{j+1}^* - x_j^*$. Bei der Betrachtung von (3.10) lässt sich erkennen, dass nun zwei unbekannte Parameter, nämlich $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)^t$ und $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^t$, geschätzt werden müssen. Wir werden jedoch zeigen, dass $\boldsymbol{\delta}$ durch $\boldsymbol{\gamma}$ ausgedrückt werden kann und somit nur noch der Schätzer von $\boldsymbol{\gamma}$ kalkuliert werden muss.

Unter Verwendung der Eigenschaften, dass der Kubischen Spline zweimal stetig differenzierbar und der Wert der zweiten Ableitung an den Stellen x_1 und x_m Null ist, kann sehr leicht bewiesen werden (siehe Wood, 2006), dass

$$\mathbf{B}\boldsymbol{\delta}^- = \mathbf{D}\boldsymbol{\gamma}.$$

Dabei ist $\boldsymbol{\delta}^- = (\delta_2, \ldots, \delta_{m-1})^t$, da $\delta_1 = \delta_m = 0$, und $\mathbf{B}$ eine $(m-2) \times (m-2)$ Matrix bzw. $\mathbf{D}$ eine $(m-2) \times m$ Matrix, welche die nachstehenden Nicht-Null Einträge besitzen:

$$\begin{aligned} B_{j,j} &= (h_j + h_{j+1})/3 & j &= 1, \ldots, m-2, \\ B_{j,j+1} &= h_{j+1}/6 & j &= 1, \ldots, m-3, \\ B_{j+1,j} &= h_{j+1}/6 & j &= 1, \ldots, m-3, \\ D_{j,j} &= 1/h_j & j &= 1, \ldots, m-2, \\ D_{j,j+1} &= -1/h_j - 1/h_{j+1} & j &= 1, \ldots, m-2, \\ D_{j,j+2} &= 1/h_{j+1} & j &= 1, \ldots, m-2. \end{aligned}$$

Mit der Definition der $(m-2) \times m$ Matrix $\mathbf{F}^- = \mathbf{B}^{-1}\mathbf{D}$ und der $m \times m$ Matrix

$$\mathbf{F} = \begin{pmatrix} \mathbf{0} \\ \mathbf{F} \\ \mathbf{0} \end{pmatrix},$$

wobei $\mathbf{0}$ ein Vektor mit lauter Nullen ist, folgt für die obige Gleichung, dass $\boldsymbol{\delta} = \mathbf{F}\boldsymbol{\gamma}$, und ergo kann (3.10) umgeschrieben werden in

$$f(x) = a_j^-(x)\gamma_j + a_j^+(x)\gamma_{j+1} + c_j^-(x)\mathbf{F}_j\boldsymbol{\gamma} + c_j^+(x)\mathbf{F}_{j+1}\boldsymbol{\gamma}, \qquad \text{falls } x_j^* \leq x \leq x_{j+1}^*. \tag{3.11}$$

Nun ist es möglich, die Kubische Spline Funktion in die von uns gewünschte Form zu bringen, nämlich

$$f(x) = \sum_{j=1}^{m} b_j(x)\gamma_j.$$

Die neuen Basisfunktionen $b_j(x)$ können von (3.11) hergeleitet werden, jedoch ist ihre genau Form nicht sofort ersichtlich. Aus diesem Grund wird im Folgenden ein Beispiel für ihre Illustration betrachtet.

Beispiel: Seien 4 Knoten gegeben, und zwar $x_1^* = 0, x_2^* = \frac{1}{3}, x_3^* = \frac{2}{3}$ und $x_4^* = 1$. Dann hat laut (3.11) die Kubische Spline Funktion für $0 \leq x \leq \frac{1}{3}$ die folgende Form inne:

$$\begin{aligned} f(x) &= a_1^-(x)\gamma_1 + a_1^+(x)\gamma_2 + c_1^-(x)\mathbf{F}_1\boldsymbol{\gamma} + c_1^+(x)\mathbf{F}_2\boldsymbol{\gamma} \\ &= \left(a_1^-(x) + c_1^-(x)\mathbf{F}_{1,1} + c_1^+(x)\mathbf{F}_{2,1}\right)\gamma_1 + \left(a_1^+(x) + c_1^-(x)\mathbf{F}_{1,2} + c_1^+(x)\mathbf{F}_{2,2}\right)\gamma_2 \\ &\quad + \left(c_1^-(x)\mathbf{F}_{1,3} + c_1^+(x)\mathbf{F}_{2,3}\right)\gamma_3 + \left(c_1^-(x)\mathbf{F}_{1,4} + c_1^+(x)\mathbf{F}_{2,4}\right)\gamma_4. \end{aligned}$$

Für die beiden übrigen Intervalle, $\left[\frac{1}{3}, \frac{2}{3}\right]$ und $\left[\frac{2}{3}, 1\right]$, kann $f(x)$ mit analoger Herangehensweise gruppiert werden. Man sieht sofort, dass nach der obigen Umsortierung die Basisfunktionen nun sehr einfach dargestellt werden können. Für $b_1(x)$ gilt z.B.:

$$\begin{aligned} b_1(x) = \mathbb{1}_{\left[0,\frac{1}{3}\right]}(x)\left[a_1^-(x) + c_1^-(x)\mathbf{F}_{1,1} + c_1^+(x)\mathbf{F}_{2,1}\right] + \mathbb{1}_{\left[\frac{1}{3},\frac{2}{3}\right]}(x)\left[c_2^-(x)\mathbf{F}_{2,1} + c_2^+(x)\mathbf{F}_{3,1}\right] \\ + \mathbb{1}_{\left[\frac{2}{3},1\right]}(x)\left[c_3^-(x)\mathbf{F}_{3,1} + c_3^+(x)\mathbf{F}_{4,1}\right]. \end{aligned}$$

Die übrigen Basisfunktionen, $b_2(x)$, $b_3(x)$ und $b_4(x)$, können analog berechnet werden. Nun sind wir daran interessiert, eine graphische Darstellung dieser Basisfunktionen zu erzeugen (siehe Abbildung 3.10).

```
> # Knoten
> m <- 4; xm <- c(0, 1:2/3, 1)

> # Definition der benötigten Funktionen und Matrizen
> h <- function(j){
+   xm[j+1]-xm[j]
+ }
> amin <- function(j,x){
+   (xm[j+1]-x)/h(j)
+ }
> aplus <- function(j,x){
+   (x-xm[j])/h(j)
+ }
> cmin <- function(j,x){
+   (((xm[j+1]-x)^3)/h(j)-h(j)*(xm[j+1]-x))/6
+ }
> cplus <- function(j,x){
```

```
+   (((x-xm[j])^3)/h(j)-h(j)*(x-xm[j]))/6
+ }
> B <- matrix(0,m-2,m-2)
> for(j in 1:(m-2)){
+   B[j,j] <- (h(j)+h(j+1))/3
+   if(j<=m-3){
+     B[j,j+1] <- h(j+1)/6
+     B[j+1,j] <- h(j+1)/6
+   }
+ }
> D <- matrix(0,m-2,m)
> for(j in 1:(m-2)){
+   D[j,j] <- 1/h(j)
+   D[j,j+1] <- -1/h(j)-1/h(j+1)
+   D[j,j+2] <- 1/h(j+1)
+ }
> Fmin <- solve(B)%*%D
> F <- rbind(rep(0,m),Fmin,rep(0,m))

> # Definition der Basisfunktion
> bj <- function(x,xm,m,j){
+   temp <- 0
+   if(j==1){
+     for(k in 1:(m-1)){
+       temp <- temp+(((x>=xm[k])*(x<xm[k+1]))*(cmin(k,x)*F[k,j]+
+                     cplus(k,x)*F[k+1,j]+(k==1)*amin(k,x)))
+     }
+     temp
+   }
+   else if(2<=j && j<=m-1){
+     for(k in 1:(m-1)){
+       temp <- temp+(((x>xm[k])*(x<=xm[k+1]))*(cmin(k,x)*F[k,j]+
+                     cplus(k,x)*F[k+1,j]+(j-1==k)*aplus(k,x)+(j==k)*amin(k,x)))
+     }
+     temp
+   }
+   else if(j==m){
+     for(k in 1:(m-1)){
+       temp <- temp+(((x>xm[k])*(x<=xm[k+1]))*(cmin(k,x)*F[k,j]+
+                     cplus(k,x)*F[k+1,j]+(k==m-1)*aplus(m-1,x)))
+     }
+     temp
+   }
+   else{
+     print("Warning: wrong j value")
```

```
+   }
+ }

> # Das Plotten der vier Basisfunktionen
> plot(xm, xm*0, ylim=c(-0.2,1), xlab="x", ylab="", cex.lab=0.8)
> title(ylab=expression(b[j](x)), mgp=c(2.0,1,0), cex.lab=0.8)
> for(j in 1:m){
+   lines(seq(0,1,length=1000), bj(seq(0,1,length=1000),xm,m,j), col=j, lty=2)
+ }
```

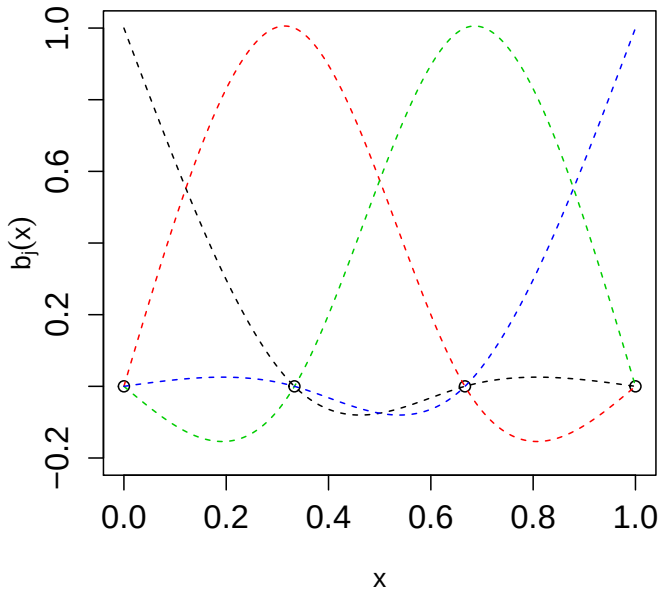

Abbildung 3.10: Die Basisfunktionen des Kubischen Splines.

Werden die Basisfunktionen nun mit den jeweiligen Koeffizienten multipliziert und im Anschluss die Resultate zusammengezählt, so ergibt sich die Kubische Spline Funktion. In Abbildung 3.11 wird eine solche Funktion dargestellt.

```
> # erstelle und plotte eine Kubische Spline Funktion
> f <- function(x,xm,m,coef){
+   temp <- 0
+   for(j in 1:m){
+     temp <- temp+(bj(x,xm,m,j)*coef[j])
+   }
+  temp
+ }
```

```
> coef <- c(0.2, -0.8, 0.6, -1.2)
> plot(xm, xm*0, ylim=c(-1.2,0.8), xlab="x", ylab="f(x)", cex.lab=0.8)
> for(j in 1:m){
+   lines(seq(0,1,length=1000), coef[j]*bj(seq(0,1,length=1000),xm,m,j),
+         col=j, lty=2)
+ }
> lines(seq(0,1,length=1000), f(seq(0,1,length=1000),xm,m,coef), lwd=2)
```

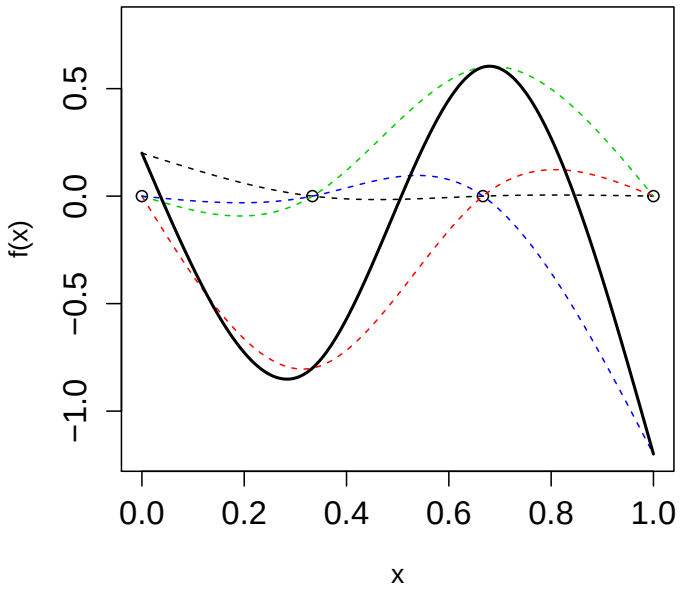

Abbildung 3.11: Eine mögliche Kubische Spline Funktion.

B-Spline bzw. P-Splines

Eine weitere Alternative für die Wahl der Basis ist der sogenannte Basis Spline, oder einfach B-Spline. Sie besitzt den Vorteil, dass die Basisfunktionen lokal sind, d.h., jede Basisfunktion ist nur auf einem bestimmten Bereich Nicht-Null. Die erste Arbeit zu B-Spline Basen geht auf Schoenberg (1946) zurück, dem folgten wichtige Werke wie z.B. Cox (1972) oder de Boor (1978). Unsere im Nachstehenden verwendeten Definitionen und Eigenschaften beruhen jedoch auf Patrikalakis und Maekawa (2002).

Eine B-Spline Funktion der Ordnung l setzt sich stückweise aus Polynomen der Ordnung $l-1$ zusammen, welche zumindest $(l-2)$-mal stetig differenzierbar sind. Um solch eine B-Spline Basis mit $m+1$ Parameter zu definieren, müssen zuerst $l+m+1$ Knoten, $x_1^* < x_2^* < \cdots < x_{l+m+1}^*$, bestimmt werden. Hinterher ist es möglich, die B-Spline Funktion

durch

$$f(x) = \sum_{j=1}^{m+1} B_j^l(x)\gamma_j, \qquad m+1 \geq l, \qquad x \in \left[x_l^*, x_{m+2}^*\right]$$

darzustellen, wobei jetzt $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{m+1})^t$ gilt und die Basisfunktionen B_j^l die folgende rekursive Definition besitzen:

$$B_j^1(x) = \begin{cases} 1 & \text{falls } x_j^* \leq x < x_{j+1}^*, \\ 0 & \text{sonst,} \end{cases}$$

für $l = 1$, und

$$B_j^l(x) = \frac{x - x_j^*}{x_{j+l-1}^* - x_j^*} B_j^{l-1}(x) + \frac{x_{j+l}^* - x}{x_{j+l}^* - x_{j+1}^*} B_{j+1}^{l-1}(x)$$

für $l > 1$ und $j = 1, \ldots, m+1$. Diese Basisfunktionen verfügen über einige wichtige Eigenschaften:

- $B_j^l(x) > 0$ für $x_j^* < x < x_{j+l}^*$.
- $B_j^l(x) = 0$ für $x_1^* \leq x \leq x_j^*$ und $x_{j+l}^* \leq x \leq x_{l+m+1}^*$.
- $\sum_{j=1}^{m+1} B_j^l(x) = 1$ für $x \in \left[x_l^*, x_{m+2}^*\right]$.

Beispiel: Um ein leichteres Verständnis der B-Splines zu ermöglichen, wird nun wieder ein Beispiel betrachtet. Sei die Ordnung der B-Spline Basis vier und die Anzahl der Parameter bzw. der Basisfunktionen sieben, also $l = 4$ und $m = 6$. Insgesamt werden daher 11 Knoten, $x_1^*, \ldots, x_{11}^*$, benötigt, welche wir der Einfachheit halber gleichmäßig auf [0,1] verteilen. Die Basisfunktionen können nun mit dem nachstehenden R-Code generiert werden, welche in Abbildung 3.12 wiedergegeben werden.

```
> # erzeuge B-Spline Basisfunktion
> Bspline <- function(x,knoten,j,l){
+   if(l==1){
+     B <- ((x>=knoten[j])*1)*((x<knoten[j+1])*1)
+   }
+   else{
+     term1 <- (x-knoten[j])/(knoten[j+l-1]-knoten[j])
+     term2 <- (knoten[j+l]-x)/(knoten[j+l]-knoten[j+1])
+     B <- term1*Bspline(x,knoten,j,l-1)+term2*Bspline(x,knoten,j+1,l-1)
+   }
+   B
+ }

> # plotte alle Basisfunktionen
```

```
> knoten <- c(0, 1:9/10, 1)
> m <- 6
> l <- 4
> plot(knoten, knoten*0, ylim=c(0,1), xlab="x", ylab="", cex.lab=0.8)
> title(ylab=expression(paste(B[j]^{l},(x))), mgp=c(2.0,1,0), cex.lab=0.8)
> for(j in 1:(m+1)){
+   lines(seq(0,1,length=1000), Bspline(seq(0,1,length=1000),knoten,j,l),
+         col=j, lty=2)
+ }
```

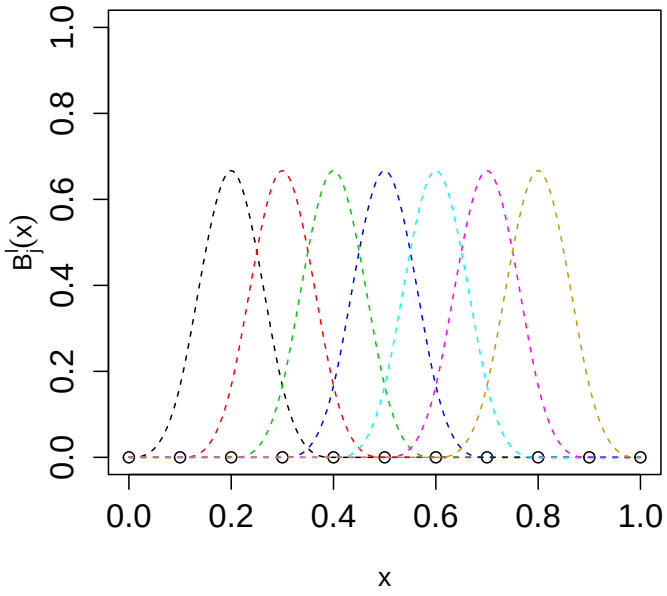

Abbildung 3.12: Die Basisfunktionen des B-Splines.

Zum Abschluss wird noch eine B-Spline Funktion, $f(x)$, erzeugt (siehe Abbilung 3.13). Zu beachten ist hier jedoch, dass das Intervall, für welches $f(x)$ bestimmt werden kann, innerhalb $[x_4^*, x_8^*]$ liegt.

```
> # definiere eine B-Spline Funktion f(x) und plotte sie mit Basisfunktionen
> f <- function(x,knoten,l,coef){
+   temp <- 0
+   if(x>=knoten[l] && x<=knoten[m+2]){
+     for(j in 1:(m+1)){
+       temp <- temp+(Bspline(x,knoten,j,l)*coef[j])
+     }
+     temp
```

```
+   }
+   else{
+     print("Warning: x value not element of domain")
+   }
+ }
> coef <- c(2.5, 2,1, -0.8, 2, 0.8, 1.6)
> plot(knoten[-c(1,2,3,9,10,11)], knoten[-c(1,2,3,9,10,11)]*0, xlab="x",
+      ylab="f(x)", xlim=c(0.3,0.7), ylim=c(-0.7,2), xaxt="n", cex.lab=0.8)
> axis(1,knoten[-c(1,2,3,9,10,11)],
+      c(expression(x[4]^{symbol("*")}),expression(x[5]^{symbol("*")}),
+        expression(x[6]^{symbol("*")}),expression(x[7]^{symbol("*")}),
+        expression(x[8]^{symbol("*")})))
> for(j in 1:(m+1)){
+   lines(seq(knoten[l],knoten[m+2],length=1000),
+         coef[j]*Bspline(seq(knoten[l],knoten[m+2],length=1000),knoten,j,l),
+         col=j, lty=2)
+ }
> lines(seq(knoten[l],knoten[m+2],length=1000),
+       f(seq(knoten[l],knoten[m+2],length=1000),knoten,l,coef), lwd=2)
```

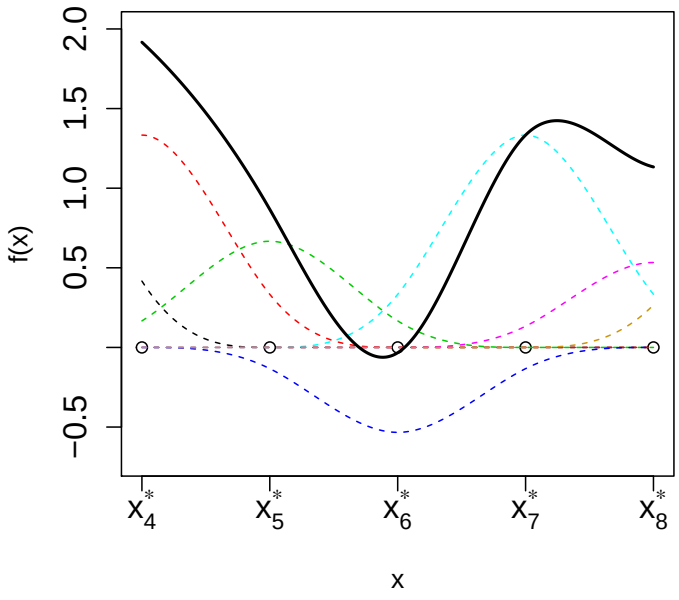

Abbildung 3.13: Eine mögliche B-Spline Funktion.

Der wahre Grund allerdings, warum B-Splines von großer Bedeutung sind, ist ihre Ver-

wendung für die sogenannten penalisierten B-Splines, oder einfach P-Splines (siehe Eilers und Marx, 1996). Werden also B-Splines als Basisfunktionen herangezogen und Differenzen-Strafterme als Strafterme bei der penalisierten Modellschätzung eingesetzt, um die Wackeligkeit der glatten Funktion zu kontrollieren, so spricht man von P-Splines. Wie schon in Abschnitt 3.1.2 kurz erwähnt, besitzt ein möglicher Differenzen-Strafterm, der Differenzen-Strafterm der Ordnung 1, die Form:

$$\mathcal{P} = \sum_{j=1}^{m} (\gamma_{j+1} - \gamma_j)^2 = \gamma_1^2 - 2\gamma_1\gamma_2 + 2\gamma_2^2 - 2\gamma_2\gamma_3 + \cdots + \gamma_{m+1}^2,$$

oder äquivalent:

$$\mathcal{P} = \boldsymbol{\gamma}^t \begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \boldsymbol{\gamma}.$$

P-Splines sind sehr beliebt, da sie sehr einfach zu konstruieren und zu verwenden sind, und eine gute Flexibilität aufweisen, d.h. Differenzen-Strafterme jeder Ordnung können mit B-Splines beliebiger Ordnung kombiniert werden. Ihr Nachteil ist jedoch, dass bei ungleichmäßiger Knotenpositionierung ihre Simplizität ein Stück weit verschwindet und die Differenzen-Strafterme verglichen mit anderen Straftermen etwas schwieriger zu interpretieren sind (für genauere Informationen siehe Wood, 2006).

3.2.2 Parameterschätzung

Es existieren im Allgemeinen einige Möglichkeiten um die Parameter eines GAMs zu schätzen, wie z.B. die in Hastie und Tibshirani (1986, 1990) erwähnte Backfitting Methode oder der Boosting Ansatz von Tutz und Binder (2006). Im Folgenden werden wir uns jedoch ausschließlich mit der in Wood (2006) angeführten Herangehensweise beschäftigen. Ist man an einem Vergleich der verschiedenen Methoden interessiert, so empfiehlt sich die Arbeit von Binder und Tutz (2008).

Erinnern wir uns nun an (3.8) zurück, also an ein GAM der Form

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\gamma},$$

wobei $\mathbf{X} = (\mathbf{X}^*|\mathbf{B}_1|\ldots|\mathbf{B}_d)$ und $\boldsymbol{\gamma} = \left(\boldsymbol{\beta}^t, \boldsymbol{\gamma}_1^t, \ldots, \boldsymbol{\gamma}_d^t\right)^t$. gilt. Wie wir bereits wissen, muss bei der Wahl der Basis auch die Anzahl der Knoten bestimmt werden, welche den Grad der Glätte der Funktion festlegen. Jedoch erweist sich das Bestimmen der optimalen Knotenzahl als eine schwierige Aufgabe, weshalb üblicherweise versucht wird, das Problem durch einen alternativen Weg zu umgehen. Dafür starten wir mit einer großen Anzahl an Basisfunktionen, q_k, welche eine ausreichende Flexibilität gewährleisten, und verringern anschließend den Grad der Glätte mit einem Strafterm. Für die Schätzung des Parameters $\boldsymbol{\gamma}$ bedeu-

tet dies, dass anstelle der Log-Likelihoodfunktion die penalisierte Log-Likelihoodfunktion

$$l_p(\boldsymbol{\gamma}|\mathbf{y}) = l(\boldsymbol{\gamma}|\mathbf{y}) - \frac{1}{2}\sum_{k=1}^{d}\lambda_k\boldsymbol{\gamma}^t\mathbf{S}_k\boldsymbol{\gamma} \tag{3.12}$$

maximiert wird, wobei der Strafterm die Zentrierungsnebenbedingung erfüllt (für genauere Informationen siehe Wood, 2006). Man erkennt, dass jetzt ausschließlich die Glättungsparameter, λ_k, die Glattheit des Modells kontrollieren und diese somit geschätzt werden müssen (siehe Abschnitt 3.2.4). Doch nun richten wir unsere Aufmerksamkeit auf die Maximierung von l_p bei gegebenen λ_k.

Die penalisierte Log-Likelihoodfunktion kann mittels der sogenannten Penalized Iteratively Reweighted Least Squares (P-IRLS) maximiert werden. Dafür wird zuerst (3.12) in

$$l_p(\boldsymbol{\gamma}|\mathbf{y}) = l(\boldsymbol{\gamma}|\mathbf{y}) - \frac{1}{2}\boldsymbol{\gamma}^t\mathbf{S}\boldsymbol{\gamma}$$

umgeschrieben, wobei $\mathbf{S} = \sum_{k=1}^{d}\lambda_k\mathbf{S}_k$ ist. Für die Scorefunktion erhalten wir unter der Verwendung von (2.4)

$$\frac{l_p(\boldsymbol{\gamma}|\mathbf{y})}{\partial\boldsymbol{\gamma}} = \frac{l(\boldsymbol{\gamma}|\mathbf{y})}{\partial\boldsymbol{\gamma}} - \mathbf{S}\boldsymbol{\gamma} = \frac{1}{\phi}\sum_{i=1}^{n}\frac{y_i-\mu_i}{a_iV(\mu_i)}\frac{\partial\mu_i}{\partial\boldsymbol{\gamma}} - \mathbf{S}\boldsymbol{\gamma}.$$

Man sieht sehr leicht, dass die Maximierung von $l_p(\boldsymbol{\gamma}|\mathbf{y})$ äquivalent zur Minimierung des penalisierten Nicht-Linearen Kleinsten Quadrate Problems

$$\mathbf{S}_p = \sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{\mathrm{Var}(y_i)} + \boldsymbol{\gamma}^t\mathbf{S}\boldsymbol{\gamma}$$

ist, falls die Terme $\mathrm{Var}(y_i)$ bekannt sind, da

$$\frac{\partial\mathbf{S}_p}{\partial\boldsymbol{\gamma}} = -2\sum_{i=1}^{n}\frac{y_i-\mu_i}{\mathrm{Var}(y_i)}\frac{\partial\mu_i}{\partial\boldsymbol{\gamma}} + 2\mathbf{S}\boldsymbol{\gamma}.$$

Nun kann leicht gezeigt werden, dass $\mathbf{S}_p$ durch

$$\mathbf{S}_p \approx \left\|\sqrt{\mathbf{W}^{(m)}}\left(\mathbf{z}^{(m)} - \mathbf{X}\boldsymbol{\gamma}\right)\right\|^2 + \boldsymbol{\gamma}^t\mathbf{S}\boldsymbol{\gamma} \tag{3.13}$$

approximiert werden kann, wobei $\mathbf{W}^{(m)}$ eine Diagonalmatrix im m-ten Iterationsschritt mit Elementen

$$w_{ii}^{(m)} = \frac{1}{a_iV(\mu_i^{(m)})}\left(\frac{\partial\mu_i^{(m)}}{\partial\eta_i^{(m)}}\right)^2$$

ist, und $\mathbf{z}^{(m)}$ einen Vektor im m-ten Iterationsschritt beschreibt, welcher die Elemente

$$z_i^{(m)} = \eta_i^{(m)} + \left(y_i - \mu_i^{(m)}\right)\left(\frac{\partial \eta_i^{(m)}}{\partial \mu_i^{(m)}}\right)$$

besitzt. Das bedeutet, dass der penalisierte Maximum-Likelihood Schätzer $\hat{\boldsymbol{\gamma}}$ durch die Iteration der folgende Schritte berechnet werden kann:

- Sei $\boldsymbol{\gamma}^{(m)}$ gegeben. Dann berechne $\boldsymbol{\mu}^{(m)}$, $\boldsymbol{\eta}^{(m)}$, $\mathbf{z}^{(m)}$ und $\mathbf{W}^{(m)}$.
- Minimiere (3.13) bezüglich $\boldsymbol{\gamma}$ um $\boldsymbol{\gamma}^{(m+1)}$ zu erhalten. Erhöhe m um eins.
- Falls die Differenz zwischen $\boldsymbol{\gamma}^{(m-1)}$ und $\boldsymbol{\gamma}^{(m)}$ signifikant klein ist, so wähle $\boldsymbol{\gamma}^{(m)}$ als penalisierten Maximum-Likelihood Schätzer.

3.2.3 Effektive Freiheitsgrade und Dispersionsparameter

Die bisher untersuchten Verfahren zur Glättung des Modells haben gemeinsam, dass zumindest ein Glättungsparameter die Glattheit der Schätzung reguliert. Bei einem Basisfunktionenansatz, also einen Ansatz ohne Strafterme, wurde die Glattheit durch die Anzahl der Basisfunktionen gesteuert, während bei einem Penalisierungsansatz ein Strafterm eingefügt wurde, welcher durch den Parameter, λ, kontrolliert wird. Es ist offensichtlich, dass ein unmittelbarer Vergleich der beiden Einflussgrößen nicht diskutabel ist, sodass ein Maß zur Bewertung der Glattheit der Modellschätzung erstrebenswert wäre. Eine leichte Möglichkeit ist ein solches Maß anhand des Linearen Modells zu begründen. Denn bei einem Linearen Modell wird die Komplexität des Modells durch die Parameteranzahl, p, wiedergegeben. Diese Anzahl steht auch mit der Spur der Hatmatrix $\mathbf{A} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$ in Verbindung, denn es gilt: $\mathrm{sp}(\mathbf{A}) = \mathrm{sp}\left(\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right) = \mathrm{sp}\left(\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{X}\right) = \mathrm{sp}\left(\mathbf{I}\right) = p$. Dies motiviert die Definition der sogenannten effektiven Freiheitsgrade (EDF) durch $\mathrm{sp}(\mathbf{A})$. Für den Fall, dass alle Glättungsparameter, λ_k, Null sind, entsprechen also die EDF der Parameteranzahl im Modell, während für $\lambda_k > 0$ die EDF an Wert verlieren und das Modell somit unflexibler wird.

Im ersten Schritt versuchen wir die EDF für ein Additives Modells zu bestimmen, also für ein Modell $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ mit Designmatrix $\mathbf{X}$, Parametervektor $\boldsymbol{\gamma}$ und $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ für jedes $i = 1, \ldots, n$. Wir wissen bereits aus Abschnitt 3.1, dass die Hatmatrix $\mathbf{A}$ die Form

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t$$

besitzt, wobei wieder $\mathbf{S} = \sum_k \lambda_k \mathbf{S}_k$ gilt. Die Spur von $\mathbf{A}$ kann nun unter der Verwendung von $\mathbf{P} = \left(\mathbf{X}^t\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t$ wie folgt umgeformt werden:

$$\mathrm{sp}(\mathbf{A}) = \mathrm{sp}\left(\mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\right) = \mathrm{sp}\left(\mathbf{XP}\right) = \sum_{j=1}^{p}\left(\mathbf{PX}\right)_{jj}.$$

Das bedeutet, dass die totalen EDF der Spur von $\mathbf{F} = \mathbf{PX} = \left(\mathbf{X}^t\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{X}$ entsprechen, während $\mathbf{F}_{jj}$ die EDF des j-ten Parameters beschreibt.

Um ein besseres Verständnis über die Bedeutung der EDF zu erhalten, ist es sinnvoll, eine abweichende Herangehensweise zu betrachten. Dafür untersuchen wir, wie stark die Strafterme die einzelnen geschätzten Parameter einschränken. Bezeichne mit

$$\tilde{\boldsymbol{\gamma}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y}$$

die Parameterschätzung des Modells ohne Strafterm, dann kann der Schätzer des Modells mit Strafterm folgendermaßen geschrieben werden:

$$\begin{aligned}\hat{\boldsymbol{\gamma}} &= \left(\mathbf{X}^t\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{y} \\ &= \left(\mathbf{X}^t\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{y} \\ &= \mathbf{F}\tilde{\boldsymbol{\gamma}}.\end{aligned}$$

$\mathbf{F}$ ist also die Matrix, welche den nicht-penalisierten Schätzer $\tilde{\boldsymbol{\gamma}}$ auf den penalisierten Schätzer $\hat{\boldsymbol{\gamma}}$ abbildet. Demzufolge ist $F_{jj} = \partial\hat{\gamma}_j/\partial\tilde{\gamma}_j$ der Schrumpffaktor des j-ten penalisierten Koeffizienten, $\hat{\gamma}_j$, oder anders gesagt misst F_{jj} wie viel sich $\hat{\gamma}_j$ ändern wird, falls $\tilde{\gamma}_j$ um eine Einheit verändert wird. F_{jj} gibt aus diesem Grund die EDF von $\hat{\gamma}_j$ wieder.

Zum Abschluss bemühen wir uns noch die EDF für ein GAM darzulegen. Unter Verwendung der Resultate von Abschnitt 3.2.2 kann sehr einfach gezeigt werden, dass der Parameterschätzer im m-ten Iterationsschritt die Form

$$\boldsymbol{\gamma}^{(m)} = \left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{z}$$

aufweist, wobei die rechte Seite im $m-1$-ten Iterationsschritt ausgewertet wurde. An dieser Stelle sei noch kurz erwähnt, dass unter Anwendung der Taylorapproximation auch leicht gezeigt werden kann, dass approximativ

$$\hat{\boldsymbol{\gamma}} \sim N(\mathbb{E}(\hat{\boldsymbol{\gamma}}), \mathbf{V}_e)$$

gilt, mit $\mathbf{V}_e = \phi\left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{X}\left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}$ und gewöhnlicherweise $\mathbb{E}(\hat{\boldsymbol{\gamma}}) \neq \boldsymbol{\gamma}$ (für genauere Informationen siehe Wood, 2006). Für die Hatmatrix, $\mathbf{A}$, ergibt sich somit die nachstehende Darstellung

$$\mathbf{A} = \mathbf{X}\left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{W}.$$

Nun ist es möglich die EDF, also die Spur von $\mathbf{A}$, zu berechnen. Unter Benutzung von $\mathbf{P} = \left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{W}$ ist dies

$$\mathrm{sp}\left(\mathbf{A}\right) = \mathrm{sp}\left(\mathbf{X}\left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{W}\right) = \mathrm{sp}\left(\mathbf{XP}\right) = \sum_{j=1}^{p}\left(\mathbf{PX}\right)_{jj}.$$

Ergo sind die totalen EDF mit der Spur von $\mathbf{F} = \mathbf{PX} = \left(\mathbf{X}^t\mathbf{W}\mathbf{X} + \mathbf{S}\right)^{-1}\mathbf{X}^t\mathbf{W}\mathbf{X}$ gleichlautend und die EDF des j-ten Parameters werden durch die Diagonalelemente von $\mathbf{F}$ gekennzeichnet.

Dispersionsparameter
Im Folgenden sind wir daran interessiert, einen Schätzer für den Dispersionsparameter, ϕ, einzuführen. Dazu betrachten wir im ersten Schritt ein AM. Für solch ein Modell entspricht die Fehlervarianz dem Dispersionsparameter, d.h. $\phi = \sigma^2$. Wir haben in Abschnitt 2.1 gesehen, dass σ^2 im Linearen Modell durch $\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2 / (n-p)$ geschätzt werden kann. Es ist nun sinnvoll, diese Herangehensweise an das AM anzupassen. Somit erhalten wir den nachstehenden Varianzschätzer

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2}{n - \mathrm{sp}(\mathbf{A})}. \tag{3.14}$$

Jedoch ist dieser Schätzer nicht erwartungstreu, da sehr einfach gezeigt werden kann, dass

$$\mathbb{E}\left(\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2\right) = \sigma^2 \left(n - 2\,\mathrm{sp}\left(\mathbf{A}\right) + \mathrm{sp}\left(\mathbf{A}^t\mathbf{A}\right)\right) + \|\boldsymbol{\mu} - \mathbf{A}\boldsymbol{\mu}\|$$

gilt.

Nachdem der additive Fall diskutiert wurde, wird nun ein Schätzer des Dispersionsparameters, ϕ, für ein GAM betrachtet. Für diesen wird meist der nachfolgende Pearson-ähnliche Schätzer herangezogen:

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

3.2.4 Wahl des Glättungsparameters

Nach all den Erläuterungen zu nichtparametrischen Regressionsverfahren muss die folgende bedeutsame und noch ausstehende Frage beantwortet werden: Wie sollten die Glättungsparameter, $\boldsymbol{\lambda}$, gewählt werden, um eine bestmögliche Beschreibung der Beobachtungen zu gewährleisten? Wir haben für ein AM in Abschnitt 3.1.3 bereits gesehen, dass $\boldsymbol{\lambda}$ mittels des Kreuzvalidierungskriteriums (CV-Kriterium) oder besser mittels des generalisierten Kreuzvalidierungskriteriums (GCV-Kriterium) geschätzt werden kann. Es wurde jedoch nicht erwähnt, dass diese Schätzungen meist verwendet werden, wenn der Dispersionsparameter unbekannt ist. Bei bekanntem ϕ empfiehlt es sich nämlich den folgenden Schätzer nach Craven und Wahba (1979) heranzuziehen.

UBRE-Kriterium
Betrachte zuerst wieder ein AM, d.h. ein Modell $\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ mit $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ für jedes $i = 1, \ldots, n$. Eine Idee für die Wahl der Glättungsparameter wäre, $\boldsymbol{\lambda}$ so zu bestimmen, dass $\hat{\boldsymbol{\mu}}$ bestmöglich den wahren Wert $\boldsymbol{\mu}$ entspricht. Ein entsprechendes Maß für diesen Abstand

ist der erwartete Mittlere Quadratische Fehler (MSE):

$$\begin{aligned}\mathbb{E}\left(\text{MSE}\right) =& \mathbb{E}\left(\frac{1}{n}\left\|\boldsymbol{\mu}-\mathbf{X}\hat{\boldsymbol{\gamma}}\right\|^2\right) = \frac{1}{n}\mathbb{E}\left(\left\|\mathbf{y}-\mathbf{A}\mathbf{y}-\boldsymbol{\epsilon}\right\|^2\right)\\ =& \frac{1}{n}\mathbb{E}\left(\left\|\mathbf{y}-\mathbf{A}\mathbf{y}\right\|^2 - 2\boldsymbol{\epsilon}^t\left(\mathbf{y}-\mathbf{A}\mathbf{y}\right)+\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}\right)\\ =& \frac{1}{n}\mathbb{E}\left(\left\|\mathbf{y}-\mathbf{A}\mathbf{y}\right\|^2 - 2\boldsymbol{\epsilon}^t\left(\boldsymbol{\mu}+\boldsymbol{\epsilon}\right)+2\boldsymbol{\epsilon}^t\mathbf{A}\left(\boldsymbol{\mu}+\boldsymbol{\epsilon}\right)+\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}\right)\\ =& \frac{1}{n}\left(\mathbb{E}\left(\left\|\mathbf{y}-\mathbf{A}\mathbf{y}\right\|^2\right) - \mathbb{E}\left(\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}\right) - 2\mathbb{E}\left(\boldsymbol{\epsilon}^t\boldsymbol{\mu}\right)+2\mathbb{E}\left(\boldsymbol{\epsilon}^t\mathbf{A}\boldsymbol{\mu}\right)+2\mathbb{E}\left(\boldsymbol{\epsilon}^t\mathbf{A}\boldsymbol{\epsilon}\right)\right).\end{aligned}$$

Wegen $\mathbb{E}\left(\boldsymbol{\epsilon}^t\boldsymbol{\epsilon}\right) = n\sigma^2$, $\mathbb{E}\left(\boldsymbol{\epsilon}^t\boldsymbol{\mu}\right) = \mathbb{E}\left(\boldsymbol{\epsilon}^t\right)\boldsymbol{\mu} = 0$, $\mathbb{E}\left(\boldsymbol{\epsilon}^t\mathbf{A}\boldsymbol{\mu}\right) = \mathbb{E}\left(\boldsymbol{\epsilon}^t\right)\mathbf{A}\boldsymbol{\mu} = 0$ und $\mathbb{E}\left(\boldsymbol{\epsilon}^t\mathbf{A}\boldsymbol{\epsilon}\right) = \mathbb{E}\left(\text{sp}\left(\boldsymbol{\epsilon}^t\mathbf{A}\boldsymbol{\epsilon}\right)\right) = \mathbb{E}\left(\text{sp}\left(\mathbf{A}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\right)\right) = \text{sp}\left(\mathbf{A}\mathbb{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^t\right)\right) = \text{sp}\left(\mathbf{A}\mathbf{I}\right)\sigma^2 = \text{sp}\left(\mathbf{A}\right)\sigma^2$ erhalten wir für die obige Gleichung

$$\mathbb{E}\left(\text{MSE}\right) = \frac{1}{n}\mathbb{E}\left(\left\|\mathbf{y}-\mathbf{A}\mathbf{y}\right\|^2\right) - \sigma^2 + \frac{2\sigma^2}{n}\text{sp}\left(\mathbf{A}\right). \tag{3.15}$$

Eine logische Konsequenz ist also, den Glättungsparameter so zu wählen, dass dieser den Schätzer des erwarteten MSE minimiert. Dies führt uns zum Un-Biased Risk Estimator- (UBRE-) Kriterium

$$V_u(\boldsymbol{\lambda}) = \frac{1}{n}\left\|\mathbf{y}-\mathbf{A}\mathbf{y}\right\|^2 - \sigma^2 + \frac{2\sigma^2}{n}\text{sp}\left(\mathbf{A}\right),$$

wobei die rechte Seite der Gleichung durch $\mathbf{A}$ von $\boldsymbol{\lambda}$ abhängt.

Nun versuchen wir das UBRE-Kriterium für ein GAM herzuleiten. Es ist möglich, die beim Fitten eines GAM verwendete Zielfunktion wie folgt niederzuschreiben

$$D(\mathbf{y},\boldsymbol{\gamma}) + \sum_k \lambda_k \boldsymbol{\gamma}^t\mathbf{S}_k\boldsymbol{\gamma}, \tag{3.16}$$

wobei der Ausdruck bezüglich $\boldsymbol{\gamma}$ minimiert wird. Ist $\boldsymbol{\lambda}$ gegeben, so haben wir in Abschnitt 3.2.2 gesehen und begründet, dass (3.16) durch

$$\left\|\sqrt{\mathbf{W}}\left(\mathbf{z}-\mathbf{X}\boldsymbol{\gamma}\right)\right\|^2 + \sum_k \lambda_k \boldsymbol{\gamma}^t\mathbf{S}_k\boldsymbol{\gamma}, \tag{3.17}$$

angenähert werden kann. Möchten wir also das UBRE-Kriterium für das Minimierungsproblem (3.16) bestimmen, so können die Argumente, welche bei der Herleitung des UBRE-Kriteriums für ein AM angewendet wurden, direkt wiederbenutzt werden und wir erhalten

$$V_u(\boldsymbol{\lambda}) = \frac{1}{n}D\left(\mathbf{y},\hat{\boldsymbol{\gamma}}\right) - \phi + \frac{2\phi}{n}\text{sp}\left(\mathbf{A}\right), \tag{3.18}$$

oder bei Verwendung der Approximation (3.17)

$$V_u^w(\boldsymbol{\lambda}) = \frac{1}{n}\left\|\sqrt{\mathbf{W}}\left(\mathbf{z}-\mathbf{X}\boldsymbol{\gamma}\right)\right\|^2 - \phi + \frac{2\phi}{n}\text{sp}\left(\mathbf{A}\right). \tag{3.19}$$

Beachte, dass V_u^w nur lokal gülitg ist, da $\mathbf{W}$ und $\mathbf{z}$ vom aktuellen $\boldsymbol{\lambda}$ abhängen.

Wir haben bisher gesehen, dass bei gegebenem Dispersionsparameter, ϕ, das Minimieren des UBRE-Kriteriums eine gute Möglichkeit ist, um den Glättungsparameter zu schätzen. Ist ϕ jedoch unbekannt, entstehen Schwierigkeiten, sodass alternative Schätzer berücksichtigt werden müssen. Ein solches Problem wird in Wood (2006) angeführt: Der Einfachheit halber betrachten wir ein AM, d.h. $\phi = \sigma^2$, und ersetzen in (3.15) $\mathbb{E}\left(\|\mathbf{y} - \mathbf{A}\mathbf{y}\|^2\right)$ durch $\sigma^2\left(n - \mathrm{sp}(\mathbf{A})\right)$, wobei die Substitution durch (3.14) gerechtfertigt wird, so gilt

$$\begin{aligned}\mathbb{E}(\mathrm{MSE}) =& \frac{1}{n}\sigma^2 n - \frac{1}{n}\sigma^2\,\mathrm{sp}(\mathbf{A}) - \sigma^2 + \frac{2\sigma^2}{n}\,\mathrm{sp}(\mathbf{A}) \\ =& \frac{\sigma^2}{n}\,\mathrm{sp}(\mathbf{A}),\end{aligned}$$

bzw. für den Schätzer des erwarteten MSE

$$\widehat{\mathbb{E}(\mathrm{MSE})} = \frac{\hat{\sigma}^2}{n}\,\mathrm{sp}(\mathbf{A}).$$

Nun veranschaulichen wir das Problem, indem wir ein Modell mit einem Parameter mit einem Modell mit zwei Parametern vergleichen. Damit das Modell mit zwei Parametern den $\widehat{\mathbb{E}(\mathrm{MSE})}$ gegenüber dem Modell mit nur einem Parameter verbessert, müsste das Modell mit zwei Parametern $\hat{\sigma}^2$ um mehr als die Hälfte reduzieren, da die Spur von $\mathbf{A}$ die EDF beschreibt. Dies kommt jedoch selten vor, sodass Modelle mit weniger Parametern bevorzugt werden. Es ist also klar, dass dieser Schätzer bei unbekanntem Dispersionsparameter nicht herangezogen werden kann.

Generalisiertes Kreuzvalidierungskriterium

Eine Alternative ist die Schätzung der Glättungsparameter bei unbekanntem Dispersionsparameter auf den Mittleren Quadratischen Vorhersagefehler (mean square prediction error: MSPE) zu stützen. Ein häufig verwendeter Ansatz, welcher für die Bestimmung des MSPE benutzt wird, ist die sogenannte Kreuzvalidierung. Dieser wurde bereits für ein AM in Abschnitt 3.1.3 eingeführt sowie begründet. Zur Wiederholung sei das resultierende Kreuzvalidierungskriterium (CV-Kriterium) nochmal angeführt:

$$\mathrm{CV}(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_i^{[-i]} - y_i\right)^2,$$

wobei $\hat{\mu}_i^{[-i]}$ wieder die Schätzung von $\mathbb{E}(y_i)$ beschreibt, falls die i-te Beobachtung y_i nicht berücksichtigt wurde. Die Verwendung des CV-Kriteriums bringt jedoch einige Probleme mit sich, denn neben dem kostspieligen Rechenaufwand hat das CV-Kriterium mit dem Fehlen der Invarianz einen störenden Nachteil inne (für genauere Informationen siehe beispielsweise Golub et al., 1979).

Aus diesem Grund wird das generalisierte Kreuzvalidierungskriterium (GCV-Kriterium) eingeführt, welche eine rotationsinvariante Form des CV-Kriteriums ist, und besitzt im Fal-

le eines AM die nachstehende Form

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{n \left\| \mathbf{y} - \hat{\boldsymbol{\mu}} \right\|^2}{\left(n - \mathrm{sp}(\mathbf{A}) \right)^2}.$$

Sind wir am GCV im generalisierten Fall interessiert, so kann dieser mit den gleichen Argumenten, welche beim UBRE-Kriterium gebraucht wurden, hergeleitet werden. Das Kriterium wird nun durch

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{nD(\mathbf{y}, \hat{\boldsymbol{\gamma}})}{\left(n - \mathrm{sp}(\mathbf{A}) \right)^2} \tag{3.20}$$

definiert. Wird anstelle der Deviance wieder die Approximation (3.17) herangezogen, dann führt dies zu

$$\mathrm{GCV}^w(\boldsymbol{\lambda}) = \frac{n \left\| \sqrt{\mathbf{W}} \left(\mathbf{z} - \mathbf{X}\boldsymbol{\gamma} \right) \right\|^2}{\left(n - \mathrm{sp}(\mathbf{A}) \right)^2}. \tag{3.21}$$

Für weitere Informationen bezüglich des GCV-Kriteriums wird angeraten, in Golub et al. (1979), Craven und Wahba (1979) oder Li (1987) nachzulesen.

Numerische Strategien zur Minimierung des UBRE- bzw. GCV-Kriteriums
Zum Abschluss dieses Abschnitts werden noch kurz zwei numerische Ansätze für die Schätzung der Glättungsparameter im Falle eines GAM bei Anwendung der UBRE- bzw. GCV-Kriterium Minimierung erklärt (für genauere Erklärungen siehe Gu und Wahba, 1991; Wood, 2006).

- Performance Iteration: In jedem Iterationsschritt des P-IRLS Schemas, die aktuellen Schätzungen seien $\boldsymbol{\gamma}^{(m)}$, $\mathbf{W}^{(m)}$ und $\mathbf{z}^{(m)}$, werden anhand des UBRE- bzw. GCV-Kriteriums die optimalen Glättungsparameter, $\boldsymbol{\lambda}^{(m)}$, kalkuliert, also $\boldsymbol{\lambda}^{(m)}$ so gewählt, dass diese die Kriterien (3.19) bzw. (3.21) minimieren. Im Anschluss wird unter Verwendung von $\boldsymbol{\lambda}^{(m)}$ der neue Iterationsschritt des P-IRLS Schemas durchgeführt, d.h. die neuen Schätzungen $\boldsymbol{\gamma}^{(m+1)}$, $\mathbf{W}^{(m+1)}$ und $\mathbf{z}^{(m+1)}$ werden mithilfe von $\boldsymbol{\lambda}^{(m)}$ berechnet und nachfolgend $\boldsymbol{\lambda}^{(m+1)}$ wieder anhand des UBRE- bzw. GCV-Kriteriums geschätzt. Dieser Prozess konvergiert leider nicht in allen Fällen, bringt aber einen Schätzer hervor, welcher numerisch effizient ist.

- Äußere Iteration (outer iteration): Aus einem fixen Satz von Glättungsparametervektoren wird für jedes Mitglied das P-IRLS Schema bis zur Konvergenz iteriert. Mit dem bei der Konvergenz erhaltenem Parametervektor wird anschließend der UBRE/GCV-Wert (siehe (3.18) bzw. (3.20)) berechnet. Mittels eines Optimierungsverfahrens wird dann der optimaler Glättungsparametervektor bestimmt. Dieser Ansatz ist im Vergleich zur Performance Iteration langsamer, jedoch ist er für Konvergenzprobleme weniger anfällig.

4 Modellierung von Staatsinsolvenzen

Für die folgenden Modellierungen wird das Softwarepaket `mgcv` von Simon N. Wood herangezogen. Ist man an einer Einführung und einer praktischen Anwendung interessiert, so empfiehlt sich Wood (2006), für detaillierte Beschreibungen der einzelnen Funktionen eignet sich hingegen Wood (2016).

In dem nachstehenden Kapitel fokussieren wir uns darauf, mithilfe von GAMs ein Modell zu finden, welches das Auftreten einer Staatsschuldenkrise in einem Schwellenland beschreibt. Dabei orientieren wir uns an den Arbeiten Manasse und Roubini (2005) und Manasse und Roubini (2009). Die beiden Ökonomen versuchen in ihrer empirischen Studie, basierend auf jährliche Beobachtungen von 47 Schwellenländern (siehe Tabelle 4.1) im Zeitraum 1970 bis 2002, die ökonomischen und politischen Merkmale eines Staatsbankrotts unter Verwendung der „Classification and Regression Tree“ (CART) Methodik zu identifizieren. Ihr Ergebnis lautet, dass 10 Merkmale einen signifikanten Einfluss auf die Zahlungsunfähigkeit eines Staates besitzen. Es wäre nun sinnvoll, diese 10 Merkmale als unsere erklärenden Variablen heranzuziehen und mit diesen ein akkurates Modell für das Eintreten einer Staatsschuldenkrise zu schätzen. Doch bevor wir uns mit der Modellfindung beschäftigen können, stellt sich die Frage, welche Daten für die Responsevariable bzw. für die 10 Prädiktoren verwendet werden sollen.

4.1 Datensatz

Für eine vernünftige Modellschätzung sind eine große Anzahl an korrekten Daten eine notwendige Voraussetzung. Jedoch sind diese in der Regel nur schwer zu bekommen und häufig auch sehr teuer. Deswegen besteht unser Interesse darin, auch um die Reproduzierbarkeit dieser Arbeit zu garantieren, Daten zu finden, welche für jeden zugänglich sind. Nach einer langen Recherche entscheiden wir uns die folgenden Quellen für die Responsevariable und für die 10 Prädiktoren zu gebrauchen:

- **Staatsschuldenkrisen** (`krisen`): Als Datensatz für die Responsevariable wird die Liste der Credit Rating Assessment Group (CRAG) von der Bank of Canada herangezogen, welcher unter der nachstehenden Adresse heruntergeladen werden kann: `www.bankofcanada.ca/wp-content/uploads/2015/05/crag-database-update-04-05-15.xlsx`. Dieser beinhaltet jährliche Informationen ab dem Jahr 1975 über die Höhe der Schulden, falls ein Land insolvent ist. Für weitere Erklärungen bzw. genauere Quellen, welche für die Berechnung der Schulden verwendet werden, empfiehlt es sich in Beers und Nadeau (2014) nachzulesen. Nun entschließen wir uns anhand der CRAG-Datenbank die binäre Responsevariable `krisen` für 47 Länder (siehe Tabelle 4.1), basierend auf die Länder in Manasse und Roubini (2005), für die Periode 1975 bis 2002 anzulegen. Diese gibt für jedes Jahr und jedes Land an, ob eine Insolvenz vorliegt oder nicht (1 = insolvent, 0 = nicht insolvent).

- **Total external debt over GDP** (`TEDY`): Wie bei der Responsevariable `krisen` bemühen wir uns nun möglichst gute Daten für die 10 erklärenden Variablen für

Argentinien (AR)	Indonesien (ID)	Pakistan (PK)
Bolivien (BO)	Israel (IL)	Polen (PL)
Brasilien (BR)	Indien (IN)	Paraguay (PY)
Chile (CL)	Jamaika (JM)	Rumänien (RO)
China (CN)	Jordanien (JO)	Russland (RU)
Kolumbien (CO)	Korea (KR)	Slowakei (SK)
Costa Rica (CR)	Kasachstan (KZ)	El Salvador (SV)
Zypern (CY)	Litauen (LT)	Thailand (TH)
Tschechien (CZ)	Lettland (LV)	Tunesien (TN)
Dominikanische Rep. (DO)	Marokko (MA)	Türkei (TR)
Algerien (DZ)	Mexiko (MX)	Trinidad und Tobago (TT)
Ecuador (EC)	Malaysia (MY)	Ukraine (UA)
Estland (EE)	Oman (OM)	Uruguay (UY)
Ägypten (EG)	Panama (PA)	Venezuela (VE)
Guatemala (GT)	Peru (PE)	Südafrika (ZA)
Ungarn (HU)	Philippinen (PH)	

Tabelle 4.1: Liste der in Manasse und Roubini (2005) verwendeten 47 Länder.

die in der Tabelle 4.1 befindlichen Länder zu finden. Die betrachtete Periode bleibt natürlich 1975 bis 2002, was 28 Jahren entspricht. Der erste Prädiktor, welcher laut Manasse und Roubini (2005) einen signifikanten Einfluss auf das Auftreten einer Staatsschuldenkrise besitzt, ist `TEDY`. Dieser lässt sich mittels WDI-Daten (World Development Indicators, World Bank) berechnen, nämlich als Quotient von „External debt stocks, total“ und „GDP“. Die WDI-Datenbank ist eine hervorragende Quelle für ökonomische und wirtschaftliche Indikatoren, da sie frei zugänglich ist und in `R` mithilfe eines Pakets leicht einzulesen ist. Für die Ermittlung von `TEDY` wird der folgende Code benutzt:

```
> # WDI package
> install.packages("WDI")
> library(WDI)

> # verwendete Schwellenländer
> threshold.country <- c("AR", "BO", "BR", "CL", "CN", "CO", "CR", "CY",
+                        "CZ", "DO", "DZ", "EC", "EG", "EE", "GT", "HU",
+                        "ID", "IN", "IL", "JM", "JO", "KZ", "KR", "LT",
+                        "LV", "MA", "MX", "MY", "OM", "PK", "PA", "PE",
+                        "PH", "PL", "PY", "RO", "RU", "SV", "SK", "TH",
+                        "TT", "TN", "TR", "UA", "UY", "VE", "ZA")
> # Anzahl Schwellenländer
> n.c <- length(threshold.country)
> # Anzahl Jahre
> n.y <- 28
```

```
> # External debt stocks, total (WDI)
> externaldebttotal <- WDI(country = threshold.country,
+                          indicator = "DT.DOD.DECT.CD",
+                          start = 1975, end = 2002, extra = FALSE,
+                          cache = NULL)
> exdetotal_WDI <- externaldebttotal$DT.DOD.DECT.CD

> # GDP
> GDP <- WDI(country = threshold.country,
+           indicator = "NY.GDP.MKTP.CD",
+           start = 1975, end = 2002, extra = FALSE, cache = NULL)
> GDP_WDI <- GDP$NY.GDP.MKTP.CD

> # Berechnung von TEDY
> TEDY <- rep(NA,n.c*n.y)
> for(i in 1:(n.c*n.y)){
+    if(!is.na(exdetotal_WDI[i]) && !is.na(GDP_WDI[i])){
+      TEDY[i] <- exdetotal_WDI[i] / GDP_WDI[i]
+    }
+ }
```

Ein Problem, welchem noch große Aufmerksamkeit gewidmet werden muss, sind die fehlenden Datenpunkte des generierten Prädiktors. Wären die verwendeten Quellen vollständig, so würde die Anzahl der für die Schätzung verfügbaren Beobachtungen 1316 betragen, da ein Zeitraum von 28 Jahren für 47 Länder betrachtet wird. Jedoch sind die von uns verwendeten Datensätze lückenhaft; bei genauerer Analyse kann festgestellt werden, dass die folgenden 17 Länder gar keine Datenpunkte aufweisen: Argentinien, Chile, Zypern, Tschechien, Estland, Ungarn, Israel, Korea, Litauen, Lettland, Oman, Polen, Russland, Slowakei, Trinidad und Tobago, Uruguay sowie Venezuela. Somit besitzt die erklärende Variable `TEDY` Informationen von 30 Ländern.

- **Short-term external debt over reserves ratio** (`STDR`): Für die Bestimmung von `STDR` werden die WDI-Daten „External debt stocks, short-term“ und „Total reserves (includes gold)“ herangezogen und deren Quotient gebildet. Auch diesmal weist der erstellte Prädiktor fehlende Datenpunkte auf, genauer gesagt besitzt `STDR` keine Informationen über dieselben 17 Länder wie der Prädiktor `TEDY`. Umgekehrt fließen also Beobachtungen von 30 Ländern in die nachfolgenden Schätzungen ein.

- **Public external debt over revenue** (`PEDR`): Den Prädiktor `PEDR` kalkulieren wir mit dem nachstehenden Ausdruck:

$$\frac{\text{Public external debt}}{(\text{Revenue (\% GDP)}/100) * \text{GDP}}.$$

Für die beiden Variablen „Public external debt“ und „GDP“ werden die WDI-Daten „External debt stocks, public and publicly guaranteed“ bzw. „GDP“ ver-

wendet. Leider kann die WDI-Datenbank für „Revenue (% GDP)“ wegen fehlender Beiträge bis zum Jahr 1990 nicht in Anspruch genommen werden. Anstelle deren kommt der Datensatz von Allan Drazen (Professor of Economics, University of Maryland) zur Anwendung, welcher auf die Daten von IFS (International Financial Statistics) beruht und unter der nachfolgenden Adresse abrufbar ist: `http://econweb.umd.edu/~drazen/Data_Sets/Data_Sets.html`. Allerdings ist auch dieser Datensatz nicht vollständig, sodass für die Länder China, Indonesien, Jamaika, Jordanien, Kasachstan, Thailand und Ukraine der Indikator „General Government Revenue (% GDP)“ der Datenbank „IMF Cross Country Macroeconomic Statistics“ eingesetzt wird. Dieser kann sehr einfach unter `https://www.quandl.com` gewonnen werden. Trotz alldem ist der generierte Prädiktor lückenhaft, denn für die folgenden 20 Länder sind keine Beobachtungen vorhanden: Argentinien, Chile, Zypern, Tschechien, Algerien, Estland, Ägypten, Ungarn, Israel, Korea, Litauen, Lettland, Marokko, Oman, Polen, Russland, Slowakei, Trinidad und Tobago, Tunesien sowie Venezuela. Für die Schätzungen der nachkommenden Abschnitte können somit nur Informationen von 27 Ländern einbezogen werden.

- **Real GDP growth** (`RGRWT`): Für den Prädiktor `RGRWT` wird die Variable „GDP growth“ der WDI-Datenbank gebraucht, welche Datenpunkte für jedes der 47 Länder innehat.

- **Inflation** (`INF`): Hierfür werden die WDI-Daten „Inflation“ herangezogen. Wie schon bei `RGRWT` besitzt `INF` Informationen von allen 47 Ländern.

- **US treasury bill rate** (`UST`): Wir verwenden für die erklärende Variable `UST`, welche für jedes Land den gleichen Wert besitzt und somit keine fehlende Einträge aufweist, die Datenbank der New York University Stern School of Business. Dieser steht unter folgender Adresse zur Verfügung: `http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html`.

- **Exchange rate overvaluation** (`EXCHRO`): Für den Prädiktor `EXCHRO` wird der Datensatz „Global Development Network Growth Database“ von William R. Easterly benutzt, welcher unter dem nachstehenden Link heruntergeladen werden kann: `http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:20701055~pagePK:64214825~piPK:64214943~theSitePK:469382,00.html`. Genauere Informationen über die Vorgehensweise zur Bestimmung dieser Variable können in Easterly und Levine (2002) nachgeschaut werden. Wie schon bei den ersten drei Prädiktoren ist der verwendete Datensatz von `EXCHRO` nicht vollständig, denn für die nachstehenden 13 Länder sind keine Datenpunkte vorhanden: China, Tschechien, Estland, Ungarn, Kasachstan, Litauen, Lettland, Oman, Polen, Rumänien, Russland, Slowakei und Ukraine. Das bedeutet also, dass für `EXCHRO` Informationen von 34 Ländern zur Verfügung stehen.

- **Exchange rate variability** (`EXCHRV`): Um `EXCHRV` zu ermitteln wird die Variable „Official exchange rate“ der WDI-Datenbank herangezogen und für diese der Varia-

tionskoeffizient über ein bewegendes Fenster der letzten 4 Jahre berechnet. Obwohl EXCHRV wie die anderen Prädiktoren nur ab 1975 betrachtet wird, muss die Variable „Official exchange rate“ ab 1972 eingelesen werden, um bei der Berechnung des Variationskoeffizienten über ein bewegendes Fenster der letzten 4 Jahre keine leeren Einträge zu erhalten. Der hierfür angewendete Code wird nachfolgend angeführt:

```
> # Official exchange rate (WDI)
> exchangerate <- WDI(country = threshold.country,
+                     indicator = "PA.NUS.FCRF",
+                     start = 1972, end = 2002, extra = FALSE,
+                     cache = NULL)
> exchangerate_WDI <- exchangerate$PA.NUS.FCRF

> # Berechnung des Variationskoeffizienten
> EXCHRV_WDI_72 <- NULL
> for(i in 1:(n.c)){
+   temp <- exchangerate_WDI[((i-1)*31+1):(i*31)]
+   temp_rev <- rev(temp)
+   cv_rev <- temp_rev
+   cv_rev[1] <- NA
+   cv_rev[2] <- NA
+   cv_rev[3] <- NA
+   for(j in 4:31){
+     cv_rev[j] <- sd(temp_rev[(j-3):j])/mean(temp_rev[(j-3):j])
+   }
+  EXCHRV_WDI_72[((i-1)*31+1):(i*31)] <- rev(cv_rev)
+ }

> # benötigen nur Werte ab 1975
> EXCHRV_WDI_matrix_all <- cbind(exchangerate[,c(1,2,4)],EXCHRV_WDI_72)
> EXCHRV_WDI_matrix<- subset(EXCHRV_WDI_matrix_all, year > 1974)
> EXCHRV <- EXCHRV_WDI_matrix$EXCHRV_WDI_72
```

Die so generierte erklärende Variable besitzt Informationen von 46 Ländern, nur von Ecuador liegen gar keine Datenpunkte vor.

- **Ratio to external financing requirements to foreign reserves** (FR): Für die Kalkulation des Prädiktors FR wird der nachstehende Ausdruck (siehe Manasse und Roubini, 2005) angewandt:

$$\frac{(\text{Account balance (\% GDP)}/100) * \text{GDP} + \text{Short term external debt}}{\text{Total reserves}}.$$

Die Daten für die in der obigen Formel benutzten Variable „Account balance (% GDP)“ werden der World Economic Outlook Database (International Monetary Fund) entnommen. Dieser Datensatz beinhaltet Werte der entsprechenden Länder ab

dem Jahr 1980 und ist unter der folgenden Adresse abrufbar: `https://www.imf.org/external/pubs/ft/weo/2010/01/weodata/index.aspx`. Für die übrigen Terme des Ausdrucks kommen die Variablen „GDP“, „External debt stocks, short-term“ bzw. „Total reserves (includes gold)“ der WDI-Datenbank zum Einsatz. Abschließend werden noch kurz die Länder angeführt, für welche der erzeugte Prädiktor keine Informationen aufweist: Argentinien, Chile, Zypern, Tschechien, Estland, Ungarn, Israel, Korea, Litauen, Lettland, Oman, Polen, Russland, Slowakei, Trinidad und Tobago, Uruguay und Venezuela. Dies entspricht 17 Ländern, d.h., dass Beobachtungen von 30 Ländern in die nachfolgenden Schätzungen eingehen.

- **Years to next presidential elections** (`YNPRE`): Das Finden geeigneter Daten für den letzten Prädiktor `YNPRE` erweist sich als besonders schwierig, sodass wir für jedes Land jede Präsidentschaftswahl separat überprüfen und die einzelnen Werte in einer Excel-Tabelle eintragen. Aufgrund nicht existierender Präsidentschaftswahlen in einigen Ländern verfügt auch dieser Prädiktor über fehlende Beobachtungen, genauer gesagt können für die nachfolgenden 19 Länder keine Einträge erstellt werden: China, Tschechien, Ägypten, Estland, Ungarn, Indonesien, Indien, Israel, Jamaika, Jordanien, Litauen, Lettland, Marokko, Malaysia, Oman, Pakistan, Thailand, Trinidad und Tobago, Türkei und Südafrika. Weitere Informationen bzw. die genauen Zahlen der Excel-Tabelle können z.B. der Homepage `https://en.wikipedia.org` entnommen werden.

4.2 Staatsschuldenkrisen über die Zeit

Nachdem wir uns im vorigen Kapitel damit beschäftigt haben, optimale Daten für unsere Responsevariable bzw. Prädiktoren zu finden, sind wir nun daran interessiert, mit diesen ein optimales Modell zu schätzen. Bevor wir uns jedoch dieser Aufgabenstellung widmen, wäre es indes sinnvoll, die Wahrscheinlichkeit des Auftretens einer Staatsschuldenkrise im Zusammenhang mit dem zeitlichen Faktor (`time`) zu kalkulieren. D.h., ein Modell zu schätzen, welches anhand eines Jahres die Ausfallwahrscheinlichkeit der Länder wiedergibt. Somit müsste es möglich sein, die globale wirtschaftliche Situation der Schwellenländer in den Jahren 1975 bis 2002 zu beschreiben und z.B. eine etwaige allgemeine Prognose für das Jahr 2003 aufzustellen.

Wie bereits erwähnt verwenden wir das Paket `mgcv` (siehe Wood, 2016) um ein GAM zu erstellen. Für unser Beispiel ergibt sich der folgende Code:

```
> library(mgcv)
> time <- rep(n.y:1, n.c)
> mod.time <- gam(krisen~s(time,bs="cr",k=10), family=binomial(link=logit))
> summary(mod.time)

Family: binomial
Link function: logit
```

```
Formula:
krisen ~ s(time, bs = "cr", k = 10)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.27342    0.06138  -4.455  8.4e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
          edf Ref.df Chi.sq p-value
s(time) 3.733  4.616  86.44  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0799   Deviance explained =  6.3%
UBRE = 0.29338  Scale est. = 1         n = 1191

> plot(mod.time, cex.lab=0.8)
```

Nach dem Laden des Pakets wird der Prädiktor `time` definiert, welcher jedes Land die Periode 1975 bis 2002 zuweist, wobei 1 für das Jahr 1975 steht und 28 für 2002. Im Anschluss wird die Funktion `gam()` aufgerufen um die Parameter des Modells zu ermitteln. Dabei beschreibt der erste Term, hier `krisen`, die binäre Responsevariable, welche als eine glatte Funktion von `time` geschätzt wird. `s()` bezeichnet die glatte Funktion. Das Argument `bs` gibt an, welche Basis benutzt wird. Wegen ihrer vorteilhaften Eigenschaft gebrauchen wir zuerst die Kubische Spline Basis und setzen deshalb `bs="cr"`. Nun muss noch die Basisdimension, `k`, bestimmt werden. Die Wahl der Basisdimension bedeutet auch, den größtmöglichen Wert, den die EDF für einen Glättungsterm innehaben kann, zu definieren. Wir versuchen es mit einem Modell mit $\texttt{k} = 10$, welches auch dem Defaultwert entspricht. Zu guter Letzt muss im `family`-Argument die Verteilung der Response und die Linkfunktion festgesetzt werden, wobei wir uns für die Binomialverteilung und den Logitlink entscheiden. Zusammenfassend wird also das folgende Modell geschätzt:

$$\mu_i = \mathbb{E}(\texttt{krisen}_i) = \pi_i = \frac{\exp(\beta_0 + \texttt{s}(\texttt{time}_i))}{1 + \exp(\beta_0 + \texttt{s}(\texttt{time}_i))},$$

wobei $\texttt{krisen}_i \sim \text{Binomial}(1, \pi_i)$ und β_0 den Intercept bezeichnet.

Im nächsten Schritt wird der `summary()` Befehl ausgeführt. Dieser beinhaltet viele nützliche Auskünfte über das Modell, wie Informationen zum parametrischen Teil (hier nur Intercept) oder zu den glatten Termen (hier nur `time`). Der p-Wert im obigen `R`-Code entspricht dem Hypothesentest, dass der Intercept bzw. die Funktion von `time` Null ist. Für den Glättungsterm ergibt sich ein p-Wert von $< 2{\cdot}10^{-16}$, d.h., dass `time` einen signifikanten Einfluss auf `krisen` besitzt. Ferner werden die effektiven Freiheitsgrade (EDF) für die glatten Terme angegeben. Der Effekt von `time` wird demnach als eine glatte Funktion mit

3.733 EDF geschätzt. Dieser Wert weist darauf hin, dass die gewählte Basisdimension von `k` = 10 in Ordnung ist. Weitere Maße für die Anpassungsgüte, die angezeigt werden, sind beispielsweise das adjustierte R^2, die erklärte Deviance, der UBRE-Wert oder der Schätzer für den Dispersionsparameter. Dabei beschreibt das adjustierte R^2 den Anteil der erklärten Varianz und ist definiert als

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \,/\, (n - \mathrm{sp}(\mathbf{A}))}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \,/\, (n-1)}.$$

Die erklärte Deviance wiederum ist der Anteil der Deviance des gefitteten Modells an der Deviance des Modells mit nur einem konstanten Term, die Null-Deviance, also gilt

$$D_{expl} = \frac{D(\mathbf{y}, \bar{y}) - D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{D(\mathbf{y}, \bar{y})} = 1 - \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{D(\mathbf{y}, \bar{y})}.$$

Mit dem letzten Befehl `plot()` wird der Graph in Abbildung 4.1 erzeugt. Dieser illustriert den geschätzten globalen Einfluss von `time` auf die Responsevariable `krisen`. Man

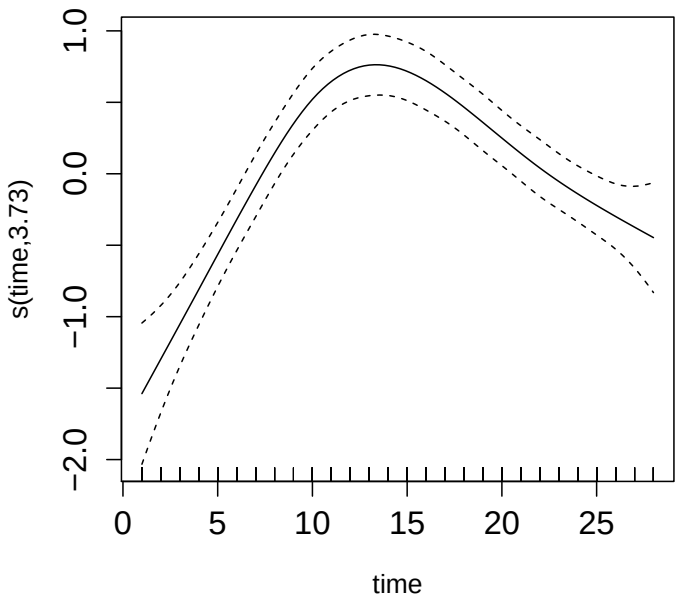

Abbildung 4.1: Der geschätzte Einfluss von `time` auf die Responsevariable `krisen`.

erkennt, dass jener zuerst linear ansteigt und ihren Höhepunkt bei zirka 13 bzw. 14, also

in den Jahren 1987 bzw. 1988 erreicht. Darauffolgend fällt er wieder linear, jedoch relativ langsam ab.

Abschließend sind wir an der geschätzten Wahrscheinlichkeit für das Eintreten einer Krise für die Jahre 1975 bis 2002 sowie an einer Vorhersage für das Jahr 2003 (`time`= 29) interessiert. Dazu betrachten wir den Graph der gefitteten Werte im Bild der empirischen relativen Häufigkeiten einer Insolvenz (siehe Abbildung 4.2).

```
> # die empirischen relativen Häufigkeiten
> zaehler <- 0
> nenner <- 0
> rel.hauf <- NULL

> for(i in 1:(n.y)){
+   for(j in 1:(n.c)){
+     if(!is.na(krisen[(j-1)*n.y+i])){
+       zaehler <- zaehler+krisen[(j-1)*n.y+i]
+       nenner <- nenner+1
+     }
+   }
+   rel.hauf[i] <- zaehler/nenner
+   zaehler<-0; nenner<-0
+ }

> # plotten der beobachteten und geschätzten Wahrscheinlichkeiten
> values <- seq(1, 29, length=5000)
> timedata <- data.frame(time=values)
> pred <- predict.gam(mod.time, timedata, se.fit=TRUE, type="response")
> plot(28:1, rel.hauf, xlab="time", ylab="Wahrscheinlichkeit einer Krise",
+      cex.lab=0.8)
> lines(values, pred$fit)
```

Man sieht, dass unser Modell die beobachtete Wahrscheinlichkeit einer Insolvenz gut wiedergibt. Darüber hinaus zeigt die Abbildung, dass die geschätzte Wahrscheinlichkeit ungefähr einen quadratischen Verlauf hat und ihren Höchstwert zwischen den Jahren 1987 und 1988 erreicht. Hierbei liegt die Wahrscheinlichkeit für eine Krise von Schwellenländern, global gesehen, bei zirka 60%. Weiters ist zu erkennen, dass die Ausfallwahrscheinlichkeit zwischen 2002 und 2003 sinkt, wobei der exakte Wert der Prognose für das Jahr 2003 noch immer 31.15% beträgt. Jedoch ist diese Vorhersage mit größter Vorsicht zu genießen, da sie global gesehen werden muss, unser Modell die Realität generell nicht allzu gut wiedergibt (betrachte hierfür den R^2_{adj}- oder D_{expl}-Wert) und eine Extrapolation prinzipiell sehr gefährlich ist (siehe Wood, 2006).

4.3 Modellfindung

Wir haben im vorigen Abschnitt gesehen, wie sich die globale Ausfallwahrscheinlichkeit der Schwellenländer über die Zeit entwickelt und haben somit einen ersten Eindruck in

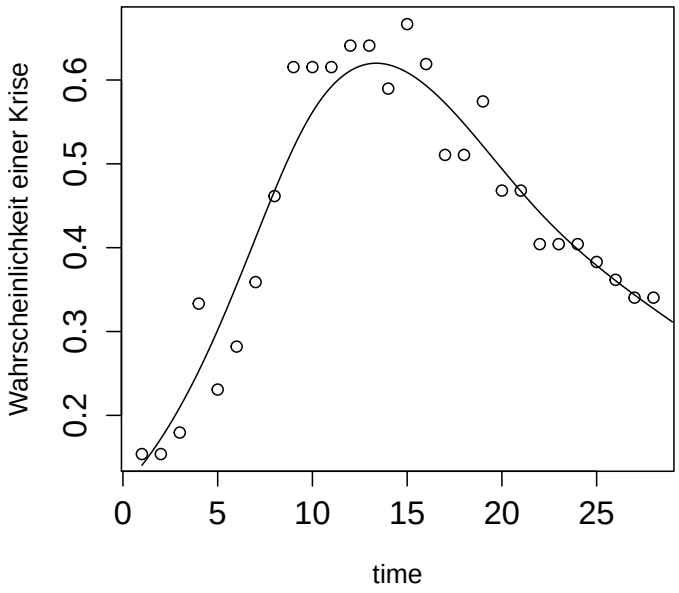

Abbildung 4.2: Beobachtete und geschätzte Wahrscheinlichkeit für das Eintreten einer Krise bezüglich `time`.

diesen Themenkomplex erhalten. Nun liegt unser Fokus darauf, ein optimales Modell mit den in Abschnitt 4.1 beschriebenen Prädiktoren zu finden. Dafür gehen wir wie folgt vor: Wir probieren schrittweise jene erklärende Variable in das Modell aufzunehmen, welcher tatsächlich signifikant ist bzw. das resultierende Modell verbessert. Es existieren einige Möglichkeiten anhand welcher der „optimale“ Prädiktor in jedem Schritt ausgewählt werden kann, wie z.B. der kleinste AIC-Wert bzw. p-Wert oder der größte R^2_{adj}-Wert. Da in unserem Modell die Beobachtungszahl für jeden Prädiktor unterschiedlich ist, kann der AIC nicht als Entscheidungsmerkmal herangezogen werden. Wir versuchen in erster Linie anhand des kleinsten p-Werts ein bestes Modell zu finden, überprüfen jedoch auch die anderen Merkmale, um die Sinnhaftigkeit unseres Modells zu gewährleisten.

4.3.1 Einzelne Betrachtung der 10 Prädiktoren

Um sich für den ersten „optimalen“ Prädiktor entscheiden zu können, wird für jede einzelne erklärende Variable ein Modell mit einem Glättungsterm betrachtet, wobei die passende Basis und Basisdimension gewählt wird. Im Anschluss werden dann die wichtigsten Merkmale für jedes Modell ermittelt und der Einfluss des Prädiktors auf `krisen` geplottet. Dieses Prozedere wird anhand von `TEDY` erklärt, für die restlichen Prädiktoren werden ausschließlich die Ergebnisse (siehe Tabelle 4.2) bzw. die Graphen (siehe Abbildungen 4.8 bis 4.16) gezeigt.

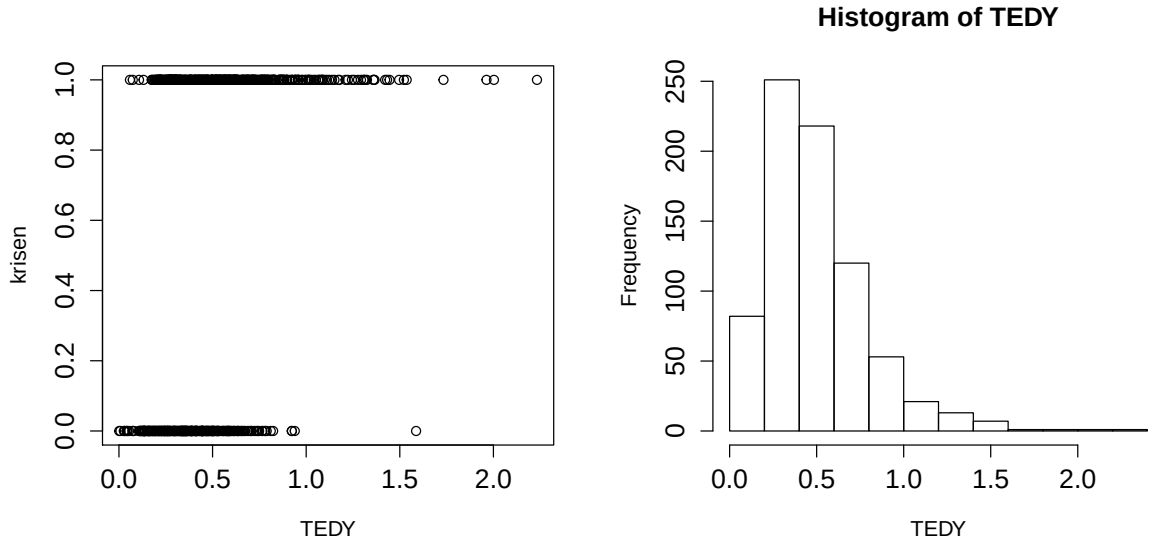

Abbildung 4.3: Links die beobachteten Werte von `TEDY` gegen `krisen`. Rechts ein Histogramm von `TEDY`.

Wie angekündigt, folgt nun die Bestimmung eines Modells mit einem Glättunsterm für den Prädiktor `TEDY`. Zuerst plotten wir die beobachteten Werte von `TEDY` gegen `krisen`, zudem erstellen wir ein Histogramm von dem Prädiktor, um auf etwaige Unregelmäßigkeiten zu prüfen. Die generierten Graphen werden in Abbildung 4.3 dargestellt. Da die Responsevariable binär ist, nimmt diese nur die Werte 0 oder 1 an, gut zu sehen im linken Plot. Auffallend ist, dass die meisten beobachteten Werte von `TEDY` im Bereich 0 bis 1 liegen. Für Werte über 1 sind kaum Beobachtungspunkte verfügbar und für die wenigen vorhandenen tritt bis auf eine Ausnahme eine Staatsschuldenkrise ein. Dies klingt sehr schlüssig, da $\texttt{TEDY} > 1$ bedeutet, dass „External debt stocks, total“ den Wert von „GDP“ übersteigt. Diese eine Ausnahme ist die Beobachtung „Indonesien 1998“, welche ein Ausreißer zu sein scheint. Denn würde dieser Datenpunkt in die Schätzung einfließen, so würde die geschätzte Wahrscheinlichkeit für das Auftreten einer Insolvenz für $\texttt{TEDY} > 1$ sehr verzerrt werden. Aus diesem Grund wird sie entfernt. Mit den übrigen Daten versuchen wir ein Modell anzupassen, wobei im ersten Ansatz eine Kubische Spline Basis mit Basisdimension $\texttt{k} = 10$ verwendet wird. Darüber hinaus wird aufgrund der binären Responsevariable natürlicherweise die Binomialverteilung angenommen, für die Linkfunktion gebrauchen wir den Logitlink.

```
> # Entfernung von Indonesien 1998
> country <- rep(threshold.country, each=n.y)
> TEDY[country=="ID" & time=="24"] <- NA

> # Modell für TEDY (Kubische Spline Basis mit Basisdimension von 10)
> mod.TEDY.1 <- gam(krisen~s(TEDY,bs="cr",k=10), family=binomial(link=logit))
> plot(mod.TEDY.1, cex.lab=0.8)
> plot(mod.TEDY.1, cex.lab=0.8, ylim=c(-6,6))
```

```
> summary(mod.TEDY.1)

Family: binomial
Link function: logit

Formula:
krisen ~ s(TEDY, bs = "cr", k = 10)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6054     0.3215   1.883   0.0597 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq  p-value
s(TEDY) 7.44  8.396  91.66 5.84e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.191   Deviance explained =   17%
UBRE = 0.15929  Scale est. = 1         n = 768
```

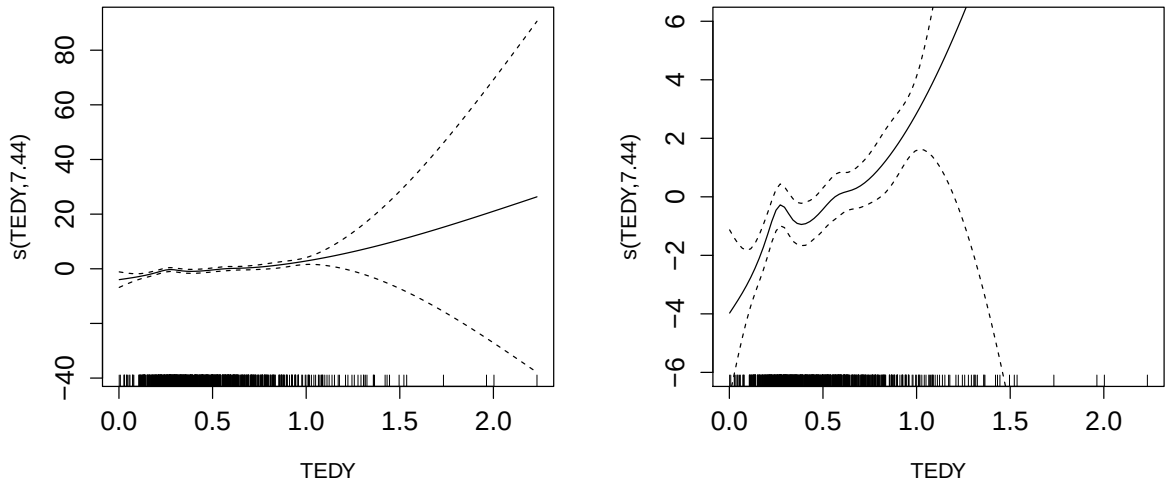

Abbildung 4.4: Der geschätzte Einfluss von `TEDY` auf die Responsevariable `krisen` mit unterschiedlicher Skalierung der y-Achse für `mod.TEDY.1`.

Bei Betrachtung des `summary()`-Outputs fällt auf, dass der Glättungsterm der erklärenden Variable `TEDY` mit einem p-Wert von $5.84 \cdot 10^{-16}$ höchst signifikant ist. Jedoch scheint

die gewählte Basisdimension zu gering zu sein, da die EDF nahe an `k` liegen. Angesichts dessen erhöhen wir die Dimension und stellen nach einigen Versuchen fest, dass eine deutliche Steigerung der EDF erst ab einer Dimension von zirka 30 erfolgt. Hier besitzen die EDF einen Wert von 19.33. Der geschätzte Glättungsterm für solch ein Modell ist aber viel zu wackelig, sodass wir bei einem Modell mit $\texttt{k} = 10$ bleiben.

Keine Aufmerksamkeit wurde bisweilen dem Term `n` des `summary()`-Outputs geschenkt. Dieser beschreibt die Anzahl der für die Schätzung verwendeten Beobachtungen, hier 768. Bei einem vollständigen Datensatz würde der Wert 1316 betragen, da ein Zeitraum von 28 Jahren für 47 Länder betrachtet wird. Wie bereits in Abschnitt 4.1 erläutert, sind die von uns verwendeten Datensätze lückenhaft, genauer gesagt sind für 17 Länder überhaupt keine Informationen vorhanden. Somit umfassen die 768 Datenpunkte Beobachtungen von 30 Ländern.

Die im obigen Code erzeugten Plots (siehe Abbildung 4.4) geben die geschätzte Wirkung von `TEDY` auf die Responsevariable wieder, wobei beim rechten Bild ausschließlich die Skalierung der y-Achse verändert wird. Man sieht, dass ein steigender, jedoch wackeliger Einfluss bis $\texttt{TEDY} < 1$ vorliegt. Für größere `TEDY`-Werte steigt die geschätzte Funktion hingegen sehr stark an und besitzt beinahe eine lineare Form.

Um etwaige Fehler im Modell zu entdecken, wird routinemäßig `gam.check()` ausgeführt. Für unser Modell erhalten die folgende Ausgabe sowie den folgenden Plot (siehe Abbildung 4.5).

```
> gam.check(mod.TEDY.1, cex.lab=0.8, cex.main=1.0)

Method: UBRE   Optimizer: outer newton
full convergence after 4 iterations.
Gradient range [1.475276e-09,1.475276e-09]
(score 0.1592936 & scale 1).
Hessian positive definite, eigenvalue range [0.0008453578,0.0008453578].
Model rank =  10 / 10

Basis dimension (k) checking results. Low p-value (k-index<1) may
indicate that k is too low, especially if edf is close to k'.

           k'   edf k-index p-value
s(TEDY) 9.000 7.440   0.932    0.02
```

Leider können die erzeugten diagnostischen Plots nur schwer interpretiert werden, da deren Betrachtung bei binären Daten im Allgemeinen nutzlos ist.

Bei unserem bisherigen Modell, `mod.TEDY.1`, wurde eine Kubische Spline Basis verwendet. Was passiert jedoch, falls eine P-Spline Basis zur Anwendung kommt?

```
> mod.TEDY.2 <- gam(krisen~s(TEDY,bs="ps",m=c(2,2),k=20),
+                   family=binomial(link=logit))
> plot(mod.TEDY.2, cex.lab=0.8)
```

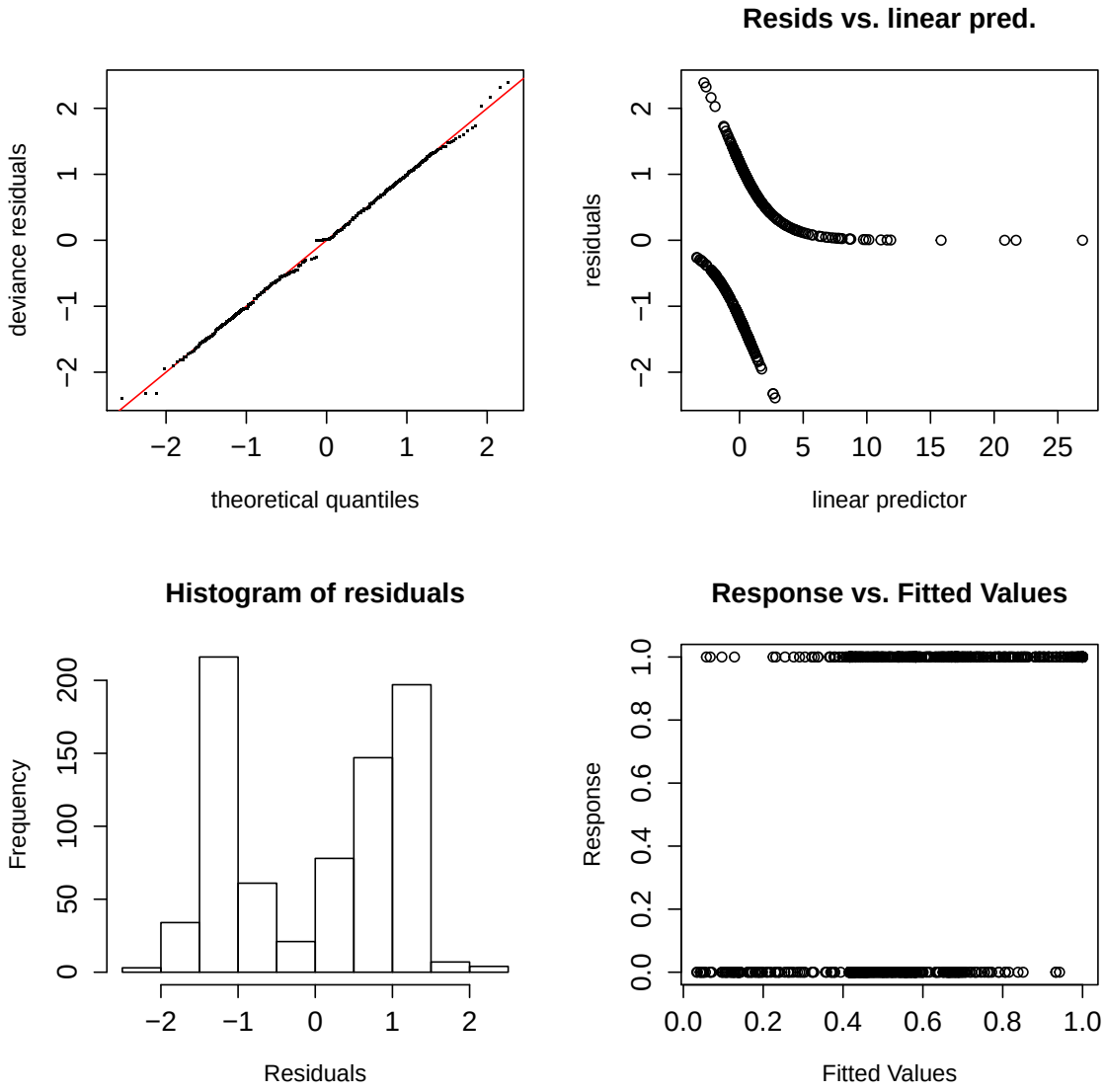

Abbildung 4.5: Diagnostische Plots für `mod.TEDY.1`.

```
> plot(mod.TEDY.2, cex.lab=0.8, ylim=c(-6,6))
> summary(mod.TEDY.2)

Family: binomial
Link function: logit

Formula:
krisen ~ s(TEDY, bs = "ps", m = c(2, 2), k = 20)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)   0.7001     0.6391   1.095    0.273

Approximate significance of smooth terms:
          edf Ref.df Chi.sq p-value
s(TEDY) 7.218  7.997  89.94 4.8e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =   0.19   Deviance explained = 16.9%
UBRE = 0.15937  Scale est. = 1         n = 768
```

Um eine P-Spline Basis einzusetzen, modifizieren wir die `gam()`-Funktion und setzen `bs= "ps"`. Zusätzlich wird ein neues Argument, `m=c(2,2)`, eingefügt, welches die genauen Spezifikationen der neuen Basis wiedergibt. Die erste Zahl des Arguments beschreibt den Grad des Splines, während die zweite Zahl die Ordnung des Differenzen-Strafterms darlegt. In unserem Beispiel entspricht `m[1]=2` dem Kubischen Spline (Achtung: Wood, 2006, benützt im Vergleich zu der in Abschnitt 3.2.1 eingeführten Version von B-Splines eine leicht abweichende Definition) und `m[2]=2` dem Differenzen-Strafterm der Ordnung 2. Nun muss noch die maximale Basisdimension festgesetzt werden, wobei wir uns nach einigen Versuchen für $k = 20$ entscheiden. Wie auch beim ersten Modell ist für uns ausschlaggebend, einen nicht allzu wackeligen Glättungsterm zu bestimmen. Die Wahl von $k = 20$ mit EDF 7.218 scheint also angemessen zu sein. Bei genauerer Betrachtung des `summary()`-Outputs sowie der geschätzten Wirkung des Prädiktors auf die Responsevariable für `mod.TEDY.2` (siehe Abbildung 4.6) stellt man fest, dass kaum Unterschiede zu unserem ersten Modell vorhanden sind. Auch der AIC-Vergleich der beiden Modelle bestätigt diese Vermutung.

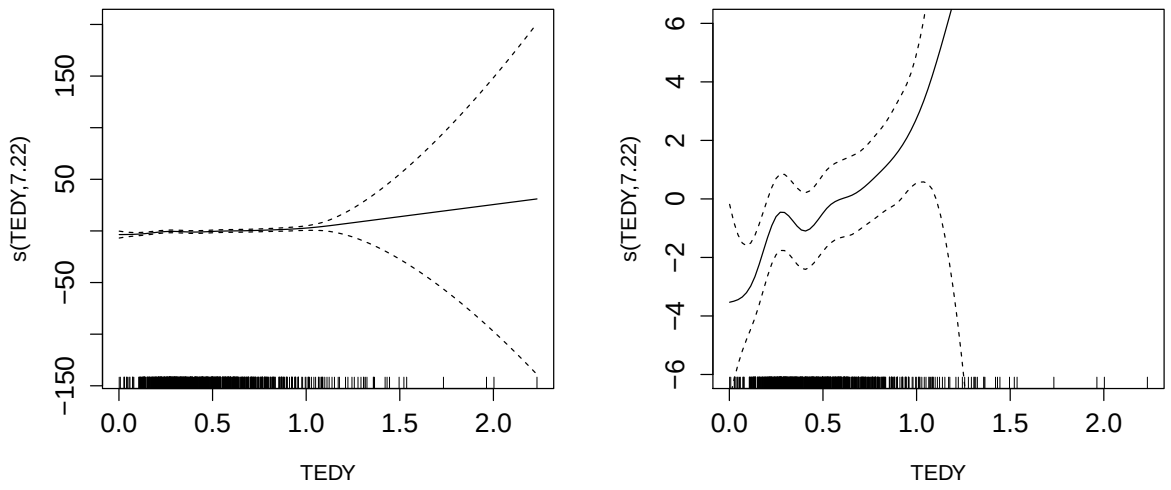

Abbildung 4.6: Der geschätzte Einfluss von `TEDY` auf die Responsevariable `krisen` mit unterschiedlicher Skalierung der y-Achse für `mod.TEDY.2`.

```
> AIC(mod.TEDY.1, mod.TEDY.2)
                 df      AIC
mod.TEDY.1 8.440261 890.3375
mod.TEDY.2 8.217813 890.3999
```

Aus diesem Grund beschließen wir `mod.TEDY.1`, also das Modell mit der Kubischer Spline Basis, dem Modell mit der P-Spline Basis vorzuziehen.

Zu guter Letzt stellt sich noch die Frage, wie die geschätzten Wahrscheinlichkeiten aussehen? Den gewünschten Graphen (siehe Abbildung 4.7) konstruieren wir mit dem folgenden Code.

```
> values <- seq(0, 2, length=5000)
> TEDYdata <- data.frame(TEDY=values)
> pred <- predict.gam(mod.TEDY.1, TEDYdata, se.fit=TRUE, type="response")
> plot(values, pred$fit, xlab="TEDY", ylab="Wahrscheinlichkeit einer Krise",
+      cex.lab=0.8, cex=0.1)
```

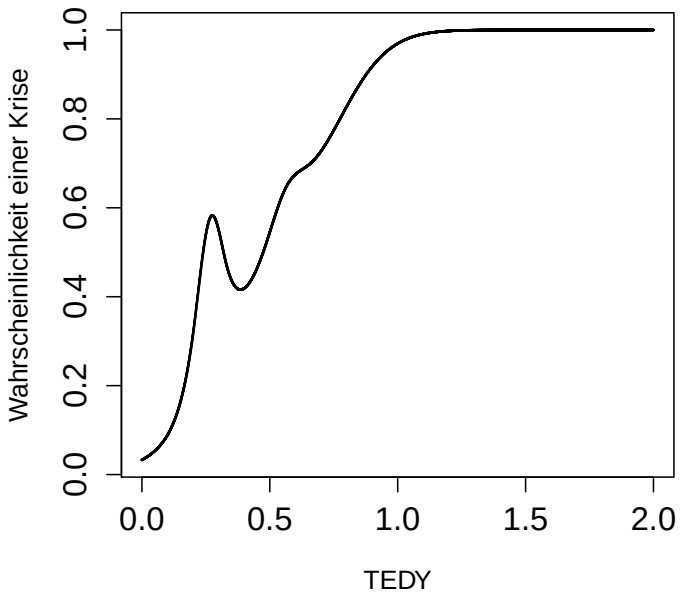

Abbildung 4.7: Die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `TEDY`.

Demzufolge steigt die Ausfallwahrscheinlichkeit der Schwellenländer bis `TEDY` $= 1$ wackelig an. Ab zirka `TEDY` > 1, d.h. falls „External debt stocks, total“ den Wert von „GDP“ übertrifft, wird mit beinahe 100%-iger Wahrscheinlichkeit eine Krise vorausgesagt. Dieser Verlauf konnte bereits in Abbildung 4.4 festgestellt werden und ist somit nicht überraschend.

Übersicht der restlichen Prädiktoren

Für die übrigen erklärenden Variablen werden nach dem obigen Schema die Glättungsterme berechnet, wobei die wichtigsten Kennzahlen der gewählten Terme in der Tabelle 4.2 wiedergegeben werden. Zu erwähnen ist hier, dass „#Länder“ die Anzahl der Länder beschreibt, welche in die Schätzung einfließen; die übrigen Kennzahlen sprechen für sich. Der geschätzte Einfluss des jeweiligen Prädiktors auf die Responsevariable sowie die jeweilige geschätzte Ausfallwahrscheinlichkeit werden in den Abbildungen 4.8 bis 4.16 dargestellt.

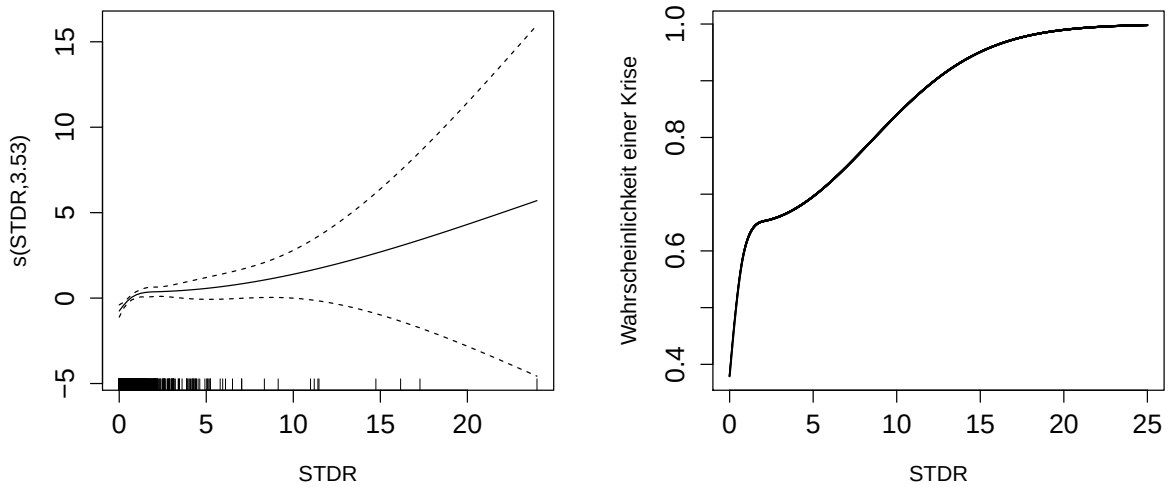

Abbildung 4.8: Links der geschätzte Einfluss von `STDR` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `STDR`.

Prädiktor	n	k	p-Wert	bs	EDF	R^2_{adj}	D_{expl}	#Länder
s(TEDY)	768	10	$5.84 \cdot 10^{-16}$	cr	7.440	0.191	17.00 %	30
s(STDR)	780	10	$3.09 \cdot 10^{-5}$	cr	3.529	0.034	2.95 %	30
s(PEDR)	517	10	$< 2 \cdot 10^{-16}$	cr	1.001	0.256	21.30 %	27
s(RGRWT)	1121	10	$3.93 \cdot 10^{-8}$	cr	3.980	0.038	3.10 %	47
s(INF)	1086	10	$8.73 \cdot 10^{-11}$	cr	3.580	0.067	5.57 %	47
s(UST)	1191	15	$5.17 \cdot 10^{-5}$	cr	8.261	0.028	2.55 %	47
s(EXCHRO)	810	10	0.0039	cr	2.462	0.017	1.57 %	34
s(EXCHRV)	1106	10	$< 2 \cdot 10^{-16}$	cr	4.533	0.090	7.07 %	46
s(FR)	643	10	0.1840	cr	3.091	0.015	1.74 %	30
s(YNPRE)	657	10	$2.02 \cdot 10^{-6}$	cr	4.834	0.065	5.50 %	28

Tabelle 4.2: Wichtige Kennzahlen der Glättungsterme für die verschiedenen Prädiktoren.

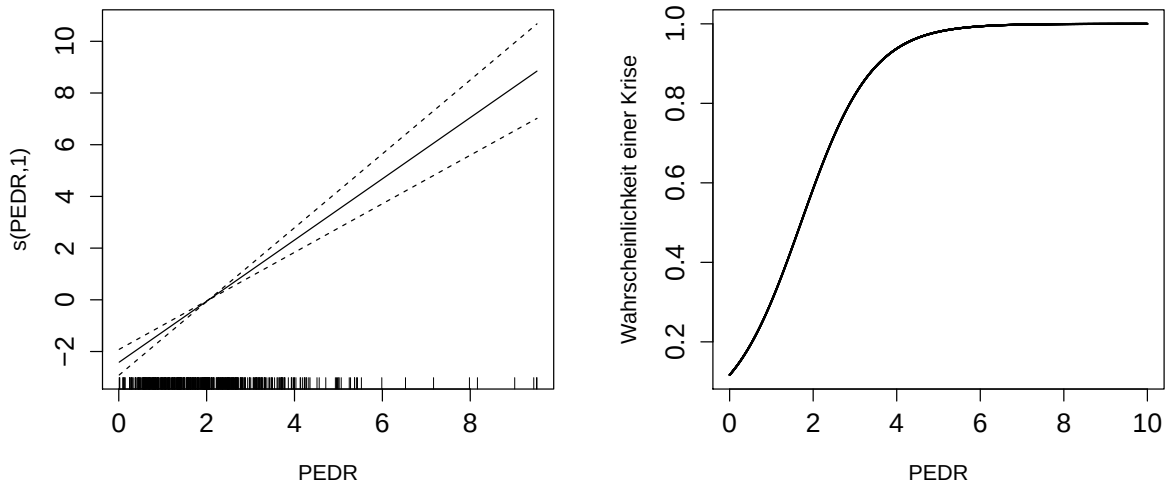

Abbildung 4.9: Links der geschätzte Einfluss von `PEDR` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `PEDR`.

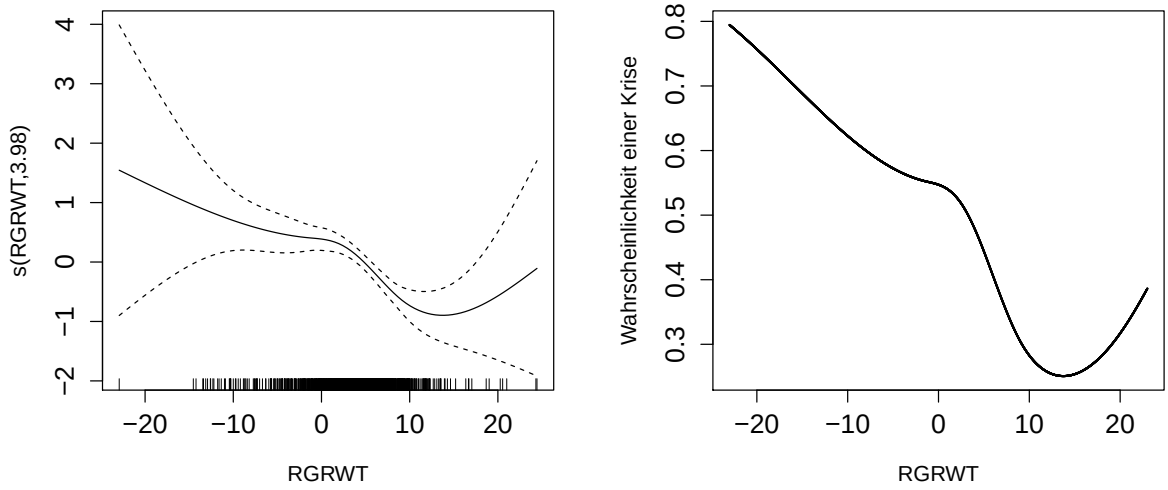

Abbildung 4.10: Links der geschätzte Einfluss von `RGRWT` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `RGRWT`.

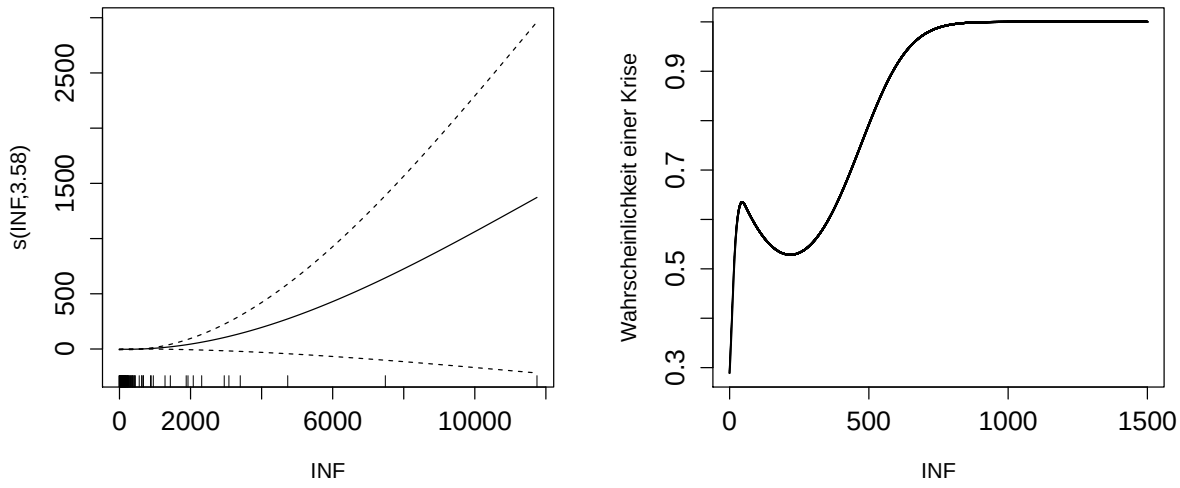

Abbildung 4.11: Links der geschätzte Einfluss von `INF` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `INF`.

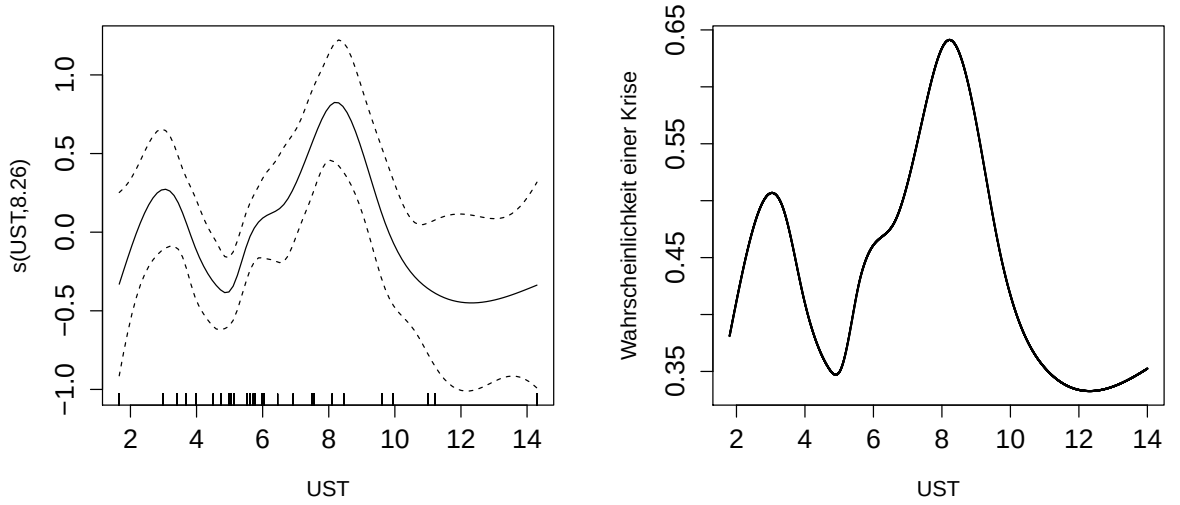

Abbildung 4.12: Links der geschätzte Einfluss von `UST` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `UST`.

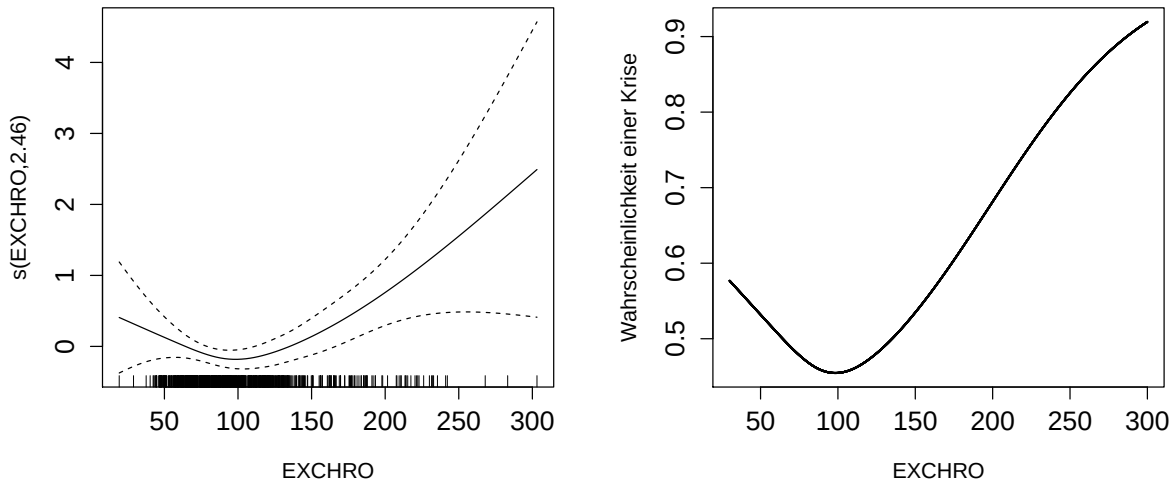

Abbildung 4.13: Links der geschätzte Einfluss von `EXCHRO` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `EXCHRO`.

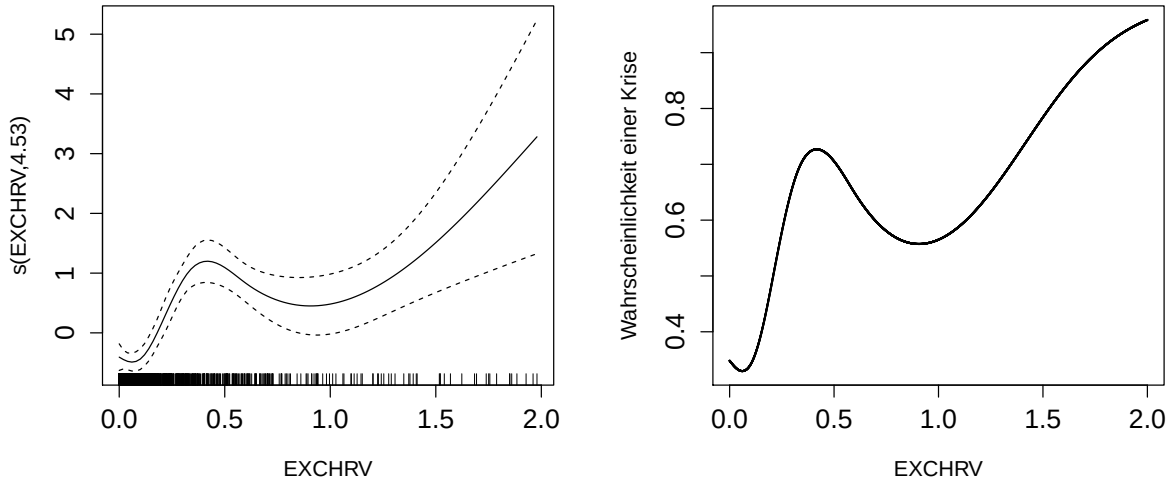

Abbildung 4.14: Links der geschätzte Einfluss von `EXCHRV` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `EXCHRV`.

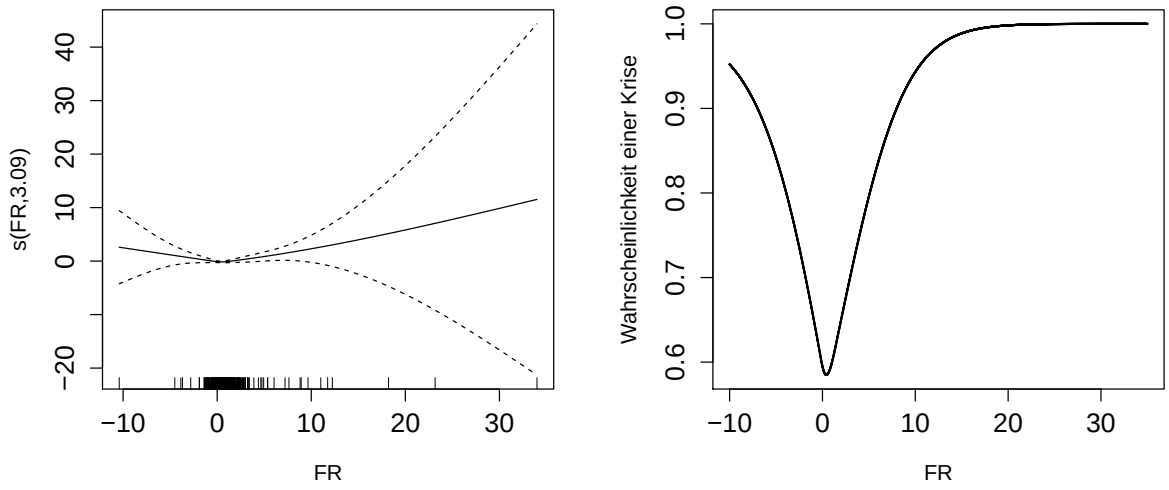

Abbildung 4.15: Links der geschätzte Einfluss von `FR` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `FR`.

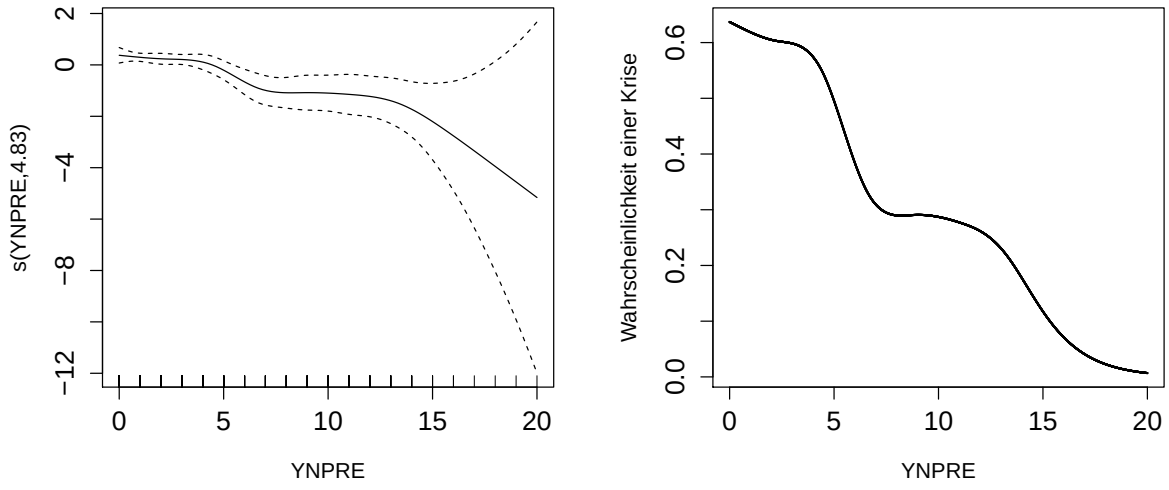

Abbildung 4.16: Links der geschätzte Einfluss von `YNPRE` auf `krisen`. Rechts die geschätzte Wahrscheinlichkeit für das Eintreten einer Krise zum jeweiligen `YNPRE`.

4.3.2 Auswahl der Prädiktoren

Wie bereits am Beginn von Abschnitt 4.3 erklärt, nehmen wir schrittweise jenen Prädiktor in das Modell auf, welche den kleinsten p-Wert innehat. Laut Tabelle 4.2 besitzen beinahe alle Prädiktoren einen signifikanten p-Wert. Nur `FR` ist nicht signifikant und wird aus diesem Grund nicht weiter betrachtet. Mit einem Wert kleiner als $2 \cdot 10^{-16}$ verfügen zwei Prädiktoren, nämlich `PEDR` und `EXCHRV`, über den kleinsten p-Wert und kommen somit in Frage. Aufgrund der höheren Anzahl von Beobachtungen bzw. Ländern, welche in die Schätzung einfließen, entscheiden wir uns die erklärende Variable `EXCHRV` in das Modell aufzunehmen. Für uns ist es von großer Wichtigkeit, Informationen von so vielen Ländern wie möglich in unser Modell einzubeziehen, um der Aufgabenstellung, Modellierung von Staatsinsolvenzen in Schwellenländern, gerecht zu werden. Nun überprüfen wir, ob der Einfluss von `EXCHRV` parametrisch oder nichtparametrisch in unser Modell eingeht. Die p-Werte der entsprechenden Tests werden in Tabelle 4.3 abgebildet. Man erkennt, dass das nicht-

Prädiktor	p-Wert (letzter Term)
`s(EXCHRV)`	$< 2 \cdot 10^{-16}$
`EXCHRV`	$7.36 \cdot 10^{-13}$
`EXCHRV`+ `EXCHRV`^{2}	0.0033
`EXCHRV`+ `EXCHRV`^{2}+ `EXCHRV`^{3}	0.0008

Tabelle 4.3: Die p-Werte des letzten Terms bei der Überprüfung des Einflusses von `EXCHRV` auf das Modell.

parametrische Modell sowie die parametrischen Modelle einen passenden p-Wert besitzen. Angesichts des kleinsten Wertes entschließen wir uns `EXCHRV` mit einem Glättungsterm in das Modell einfließen zu lassen.

Insgesamt bleiben jetzt noch acht Prädiktoren übrig, die zusätzlich zu `s(EXCHRV)` eingefügt werden können. Tabelle 4.4 gibt einige wichtige Kennzahlen wieder, falls deren Einfluss mit einem Glättungsterm in das Modell eingeht. Demnach hat jede erklärende

Prädiktor	n	`k`	p-Wert	`bs`	EDF	R^2_{adj}	D_{expl}	#Länder
+`s(TEDY)`	731	15	$1.80 \cdot 10^{-11}$	`cr`	8.959	0.262	24.60 %	29
+`s(STDR)`	744	10	0.0036	`cr`	3.568	0.077	7.02 %	29
+`s(PEDR)`	486	10	$< 2 \cdot 10^{-16}$	`cr`	1.000	0.283	24.10 %	26
+`s(RGRWT)`	1068	10	0.0003	`cr`	3.811	0.108	8.80 %	46
+`s(INF)`	1037	10	0.0060	`cr`	4.655	0.101	8.54 %	46
+`s(UST)`	1106	10	0.0019	`cr`	6.573	0.108	8.83 %	46
+`s(EXCHRO)`	785	10	0.0012	`cr`	2.090	0.102	8.47 %	33
+`s(YNPRE)`	588	10	0.0006	`cr`	6.527	0.126	11.50 %	27

Tabelle 4.4: Wichtige Kennzahlen der Glättungsterme für die verschiedenen Prädiktoren, falls diese zusätzlich zu `s(EXCHRV)` ins Modell eingeführt werden.

Variable einen signifikanten p-Wert inne, sodass alle acht Prädiktoren als Kandidaten in

Betracht kommen. Den kleinsten p-Wert besitzt PEDR, unsere Wahl fällt jedoch auf den Prädiktor mit dem zweitkleinsten p-Wert, nämlich TEDY. Die Entscheidung wird abermals damit begründet, dass bei einem Modell mit TEDY verglichen mit PEDR die Anzahl der in die Schätzungen eingehenden Beobachtungen mit 731 um einiges höher ist. Darüber hinaus werden Informationen von einer größeren Anzahl von Ländern für die Schätzung berücksichtigt, hingegen können kaum Unterschiede beim R^2_{adj}- bzw. D_{expl}-Wert wahrgenommen werden. Nun stellt sich wieder die Frage, ob der neue Prädiktor parametrisch oder nichtparametrisch ins Modell aufgenommen werden soll. Tabelle 4.5 legt zwar einen linearen

Prädiktor	p-Wert (letzter Term)
s(EXCHRV)+s(TEDY)	$1.80 \cdot 10^{-11}$
s(EXCHRV)+TEDY	$< 2 \cdot 10^{-16}$[a]
s(EXCHRV)+TEDY+TEDY^2	0.3598[a]
s(EXCHRV)+TEDY+TEDY^2+TEDY^3	0.0003[a]

[a] Modell ist zwar signifikant, jedoch verringert sich der D_{expl}-Wert um mehr als 20%.

Tabelle 4.5: Die p-Werte des letzten Terms bei der Überprüfung des Einflusses von TEDY auf das bisherige Modell.

Einfluss nahe, allerdings reduziert sich der D_{expl}-Wert um mehr als 20%, sodass wir TEDY mit einem Glättungsterm einfließen lassen.

In dem nächsten Schritt werden für die restlichen sieben Prädiktoren, welche noch ergänzend aufgenommen werden können, die wichtigsten Kennzahlen ermittelt (siehe Tabelle 4.6). Demzufolge besitzen die Prädiktoren STDR, INF, UST und YNPRE keinen signifi-

Prädiktor	n	k	p-Wert	bs	EDF	R^2_{adj}	D_{expl}	#Länder
+s(STDR)	731	10	0.3730	cr	7.291	0.269	25.90 %	29
+s(PEDR)	485	10	$3.96 \cdot 10^{-9}$	cr	1.137	0.362	35.20 %	25
+s(RGRWT)	726	10	0.0456	cr	4.391	0.271	25.70 %	29
+s(INF)	706	10	0.0721	cr	3.611	0.265	25.00 %	29
+s(UST)	731	10	0.5590	cr	1.000	0.261	24.70 %	29
+s(EXCHRO)	566	10	$2.53 \cdot 10^{-6}$	cr	4.584	0.323	30.10 %	25
+s(YNPRE)	385	10	0.0599	cr	5.097	0.186	19.90 %	17

Tabelle 4.6: Wichtige Kennzahlen der Glättungsterme für die verschiedenen Prädiktoren, falls diese zusätzlich zu s(EXCHRV) und s(TEDY) ins Modell eingeführt werden.

kanten p-Wert und werden nicht weiter betrachtet. Übrig bleiben PEDR, RGRWT und EXCHRO, wir entschließen uns jedoch keinen weiteren Prädiktor in das Modell eingehen zu lassen. Dies hat die folgenden Gründe: Das Hinzufügen von RGRWT würde das Modell kaum verbessern (siehe R^2_{adj}- bzw. D_{expl}-Wert), sondern nur unnötig verkomplizieren. Dagegen würde die Aufnahme von PEDR bzw. EXCHRO die Anzahl der für die Schätzung verwendeten Beobachtungen deutlich reduzieren, sodass wichtige Informationen verloren gehen könnten. Also werden in diesem Schritt keine weiteren Prädiktoren eingefügt und wir haben somit ein mögliches optimales Modell gefunden.

```
> mod.krisen <- gam(krisen~s(EXCHRV,bs="cr",k=10)+s(TEDY,bs="cr",k=15),
                    family=binomial(link=logit))
```

4.3.3 Neuer Prädiktor: Total reserves (TORES)

Im letzten Abschnitt wurde ein mögliches optimales Modell für das Eintreten einer Staatsschuldenkrise bestimmt, welches einen Glättungsterm für EXCHRV und TEDY beinhaltet. Überraschend ist jedoch, dass die erklärende Variable STDR, das Verhältnis der kurzfristigen Auslandsverschuldung zu den Reserven eines Landes, nicht für das Modell berücksichtigt wurde. Es verwundert ein wenig, dass die Reserven eines Landes in keiner Art und Weise in das Modell einfließen und somit nicht für das Auftreten einer Krise verantwortlich sind. Aus diesem Grund untersuchen wir, ob anstelle von STDR möglicherweise die WDI-Variable „Total reserves (includes gold)“ (TORES) einen signifikanten Einfluss aufweist und das Modell verbessert. Dafür fügen wir einen Glättungsterm für TORES den zwei bereits ausgewählten Prädiktoren hinzu und schätzen das neue Modell.

```
> mod.TORES <- gam(krisen~s(EXCHRV,bs="cr",k=10)+s(TEDY,bs="cr",k=15)+
+                  s(TORES,bs="cr",k=20), family=binomial(link=logit))
> summary(mod.TORES)

Family: binomial
Link function: logit

Formula:
krisen ~ s(EXCHRV, bs = "cr", k = 10) + s(TEDY, bs = "cr", k = 15) +
    s(TORES, bs = "cr", k = 20)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.7763     1.4108    0.55    0.582

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(EXCHRV)  5.622  6.314  60.13 8.79e-11 ***
s(TEDY)   11.927 13.132  58.48 1.13e-07 ***
s(TORES)  15.950 17.014 109.31 1.67e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.453   Deviance explained = 41.9%
UBRE = -0.10666  Scale est. = 1         n = 731

> AIC(mod.TORES, mod.krisen)
                  df      AIC
mod.TORES  34.49914 653.0289
mod.krisen 14.14223 785.2953
```

Und tatsächlich, die Hinzunahme von TORES ist dem p-Wert zufolge adäquat und verbessert das Modell überdies erheblich. Der R^2_{adj}-Wert sowie der D_{expl}-Wert erhöhen sich auf 45.3% bzw. 41.9%. Da die Beobachtungszahl keiner Änderung unterliegt, können die AIC-Werte miteinander verglichen werden und wir registrieren eine deutliche Verringerung von 785.30 auf 653.03. Darüber hinaus bleibt auch die Anzahl der Länder, welche in die Schätzung einfließen, mit 29 unverändert. Somit spricht also nichts dagegen, TORES als nichtparametrischen Einfluss in das Modell aufzunehmen. Trotzdem werden noch parametrische Modelle für TORES getestet, deren p-Werte in der Tabelle 4.7 angezeigt werden. Man sieht, dass die

Prädiktor	p-Wert (letzter Term)
s(EXCHRV)+s(TEDY)+s(TORES)	$1.67 \cdot 10^{-15}$
s(EXCHRV)+s(TEDY)+TORES	$3.39 \cdot 10^{-10}$[a]
s(EXCHRV)+s(TEDY)+TORES+TORES^2	0.0225[a]
s(EXCHRV)+s(TEDY)+TORES+TORES^2+TORES^3	0.0042[a]

[a] Modell ist zwar signifikant, jedoch verringert sich der D_{expl}-Wert um mehr als 20%.

Tabelle 4.7: Die p-Werte des letzten Terms bei der Überprüfung des Einflusses von TORES auf das bisherige Modell.

parametrischen Modelle zwar signifikant sind, jedoch reduziert sich ihr D_{expl}-Wert um mehr als 20%. Vor diesem Hintergrund und auch aufgrund des kleinsten p-Werts entscheiden wir uns den neuen Prädiktor mit einem Glättungsterm in das Modell einfließen zu lassen.

4.3.4 Länderspezifischer Faktor

Wir haben mit mod.TORES ein mögliches Modell für die Ausfallwahrscheinlichkeit von Schwellenländern gefunden, jedoch nicht berücksichtigt, dass diese auch vom betrachteten Land selbst abhängen könnte. Es wäre also sinnvoll, zusätzlich einen länderspezifischen Faktor, f_country, dem Modell hinzuzufügen.

Im Gegensatz zu den vorigen Abschnitten gebrauchen wir nun für die Erstellung des Modells ausnahmsweise die Funktion bam() von Wood (2016), welche speziell bei großen Datenmengen verwendet wird. Wir wollen damit zeigen, dass auch eine alternative Funktion des Pakets mgcv zur Schätzung von GAMs herangezogen werden kann.

```
> country <- rep(threshold.country, each=n.y)
> f_country <- as.factor(country)
> mod.TORES.factor <- bam(krisen~s(EXCHRV,bs="cr",k=10)+s(TEDY,bs="cr",k=15)+
+                         s(TORES,bs="cr",k=20)+f_country,
+                         family=binomial(link=logit), method="GCV.Cp")
> anova(mod.TORES.factor)

Family: binomial
Link function: logit

Formula:
```

```
krisen ~ s(EXCHRV, bs = "cr", k = 10) + s(TEDY, bs = "cr", k = 15) +
    s(TORES, bs = "cr", k = 20) + f_country

Parametric Terms:
          df Chi.sq p-value
f_country 28  69.33 2.3e-05

Approximate significance of smooth terms:
             edf Ref.df Chi.sq  p-value
s(EXCHRV)  6.753  7.408  28.82 0.000233
s(TEDY)    6.314  7.837  40.95 2.04e-06
s(TORES)  15.056 16.217  37.15 0.002241
```

Um den Einfluss eines Faktors zu überprüfen, kommt die Funktion `anova()` zur Anwendung. Man sieht, dass durch die Hinzunahme von `f_country` die drei Prädiktoren signifikant bleiben, als auch der länderspezifische Faktor einen signifikanten p-Wert von $2.3 \cdot 10^{-5}$ aufweist. Des weiteren werden für `f_country` nachvollziehbare Freiheitsgrade (`df`) ermittelt. Der kalkulierte `df`-Wert von 28 bedeutet, dass 29 Faktorstufen erzeugt werden; also wird tatsächlich für jedes Land, welche in die Schätzung eingeht, ein Parameter generiert. In der Konsequenz bedeutet dies, dass die Verwendung eines länderspezifischen Faktors in Betracht gezogen werden sollte.

Bei genauerer Analyse stellt sich jedoch heraus, dass die Aufnahme von `f_country` in das Modell einige Probleme mit sich bringt, welche im Folgenden diskutiert werden: Für acht Länder, genauer gesagt für China (CN), Dominikanische Rep. (DO), Ägypten (EG), Jordanien (JO), Malaysia (MY), Thailand (TH), Ukraine (UA) und Südafrika (ZA), liegt im betrachteten Zeitraum „nie“ oder „immer“ eine Krise vor, sodass die länderspezifischen Parameter dieser Staaten numerisch durch $-\infty$ bzw. ∞ geschätzt werden. Ein geschätzter Parameter von $-\infty$ generiert ein $\hat{\pi} = 0$, also die geschätzte Ausfallwahrscheinlichkeit eines Schwellenlandes von Null, im Gegenzug ist ein Schätzer von ∞ mit $\hat{\pi} = 1$ gleichzusetzen. Dies impliziert, dass für die obigen acht Länder unabhängig vom Wert der drei Prädiktoren mit einer Wahrscheinlichkeit von Eins entweder eine krisenfreie Zeit oder umgekehrt eine krisenbehaftete Zeit prognostiziert wird. Ein Modell, bei welchem durch die Modifikation der Prädiktorwerte für einige Länder keine Veränderung der Eintrittswahrscheinlichkeit einer Krise erfolgt, macht natürlich wenig Sinn, sodass wir uns entscheiden, den länderspezifischen Faktor, `f_country`, doch nicht in unser Modell einfließen zu lassen.

Insgesamt kann also festgehalten werden, dass bei unserem Modell die Ausfallwahrscheinlichkeit nicht vom betrachteten Land abhängt, sondern ausschließlich von den Prädiktoren `EXCHRV`, `TEDY` und `TORES`. Wir haben somit unser finales Modell, `mod.TORES`, gefunden und wir setzen:

```
> mod.final <- mod.TORES
```

Für den linearen Prädiktor ergibt sich schlussendlich der nachstehende Term, wobei die geschätzten Funktionen für `EXCHRV`, `TEDY` und `TORES` in der Abbildung 4.17 dargestellt werden:

$$\hat{\eta} = \hat{f}_1(\texttt{TEDY}) + \hat{f}_2(\texttt{EXCHRV}) + \hat{f}_3(\texttt{TORES}) + 0.7763,$$

und folglich für die geschätzte Ausfallwahrscheinlichkeit eines Schwellenlandes:

$$\hat{\pi} = \frac{\exp(\hat{\eta})}{1 + \exp(\hat{\eta})}.$$

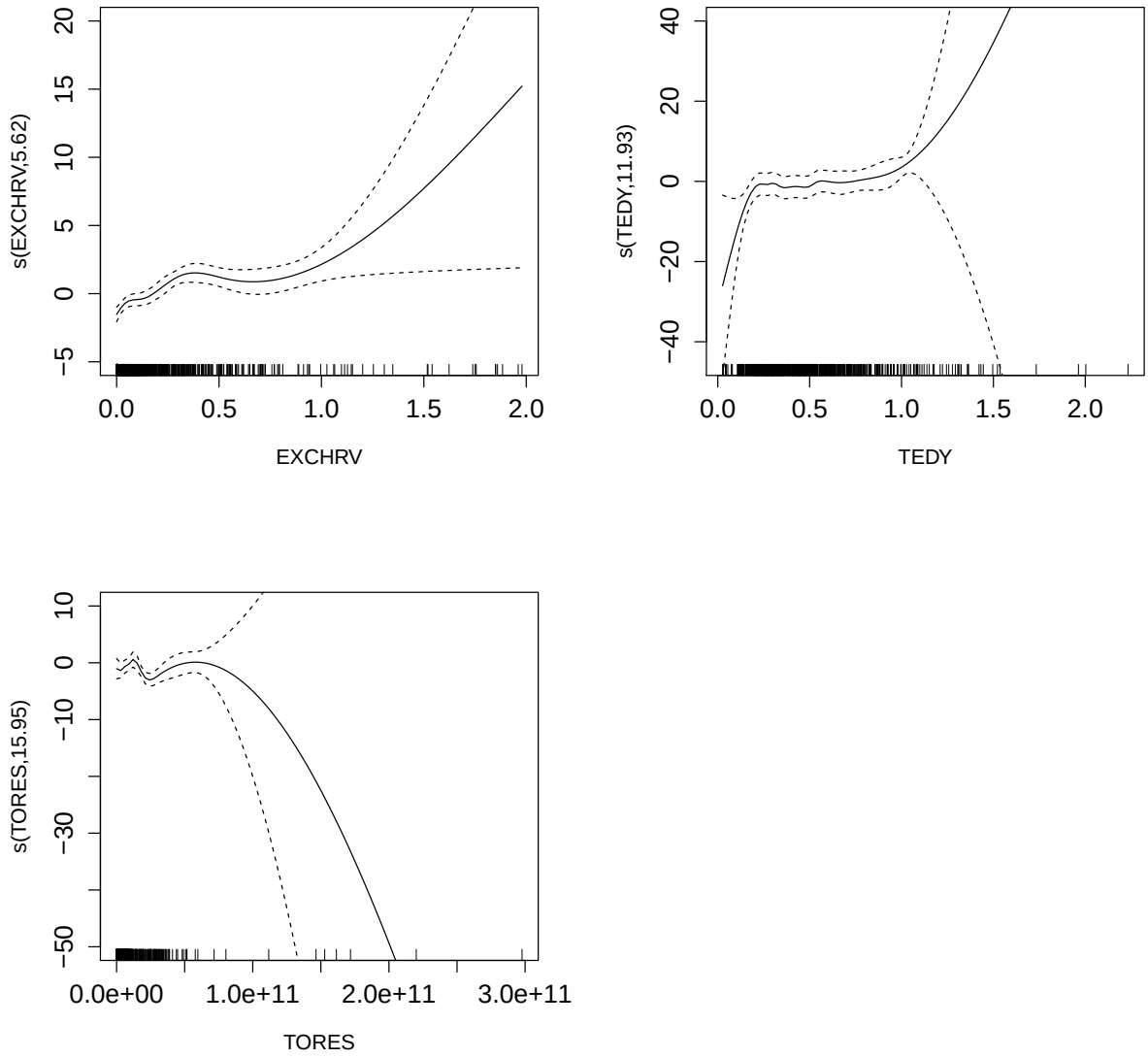

Abbildung 4.17: Der geschätzte Einfluss von `EXCHRV`, `TEDY` bzw. `TORES` auf `krisen` unter dem finalen Modell.

4.4 Vorhersage

Um die Wahrscheinlichkeit einer Staatsschuldenkrise für das Jahr 2003 vorherzusagen, benötigt unser Modell die Werte der drei Prädiktoren, welche im Jahr 2002 natürlich noch nicht vorhanden sind. Es wäre demnach vernünftig, diese für das Jahr 2003 zu schätzen und anschließend in unser Modell einzusetzen, um somit die Insolvenz eines Landes vorauszusagen. Nachfolgend ist es dann auch möglich, die Prognose mit den tatsächlichen Werten zu vergleichen und so den Vorhersagefehler zu bestimmen.

4.4.1 Vorhersage der Prädiktoren

Im ersten Schritt versuchen wir also den Vorhersagewert der erklärenden Variablen für das Jahr 2003 zu ermitteln. Dafür eruieren wir mithilfe von GAMs ein Modell, welches den Erwartungswert des jeweiligen Prädiktors bezüglich der Zeit (`time`) schätzt. Wir überprüfen zudem, ob der Erwartungswert zusätzlich vom betrachteten Land abhängt. Mit dem so konstruierten Modell können wir dann eine Prognose für die entsprechende erklärende Variable erstellen.

Wir beginnen mit dem Prädiktor `TEDY`. Bevor nun ein Modell generiert werden kann, muss zuallererst die Verteilung der erklärenden Variable bestimmt werden. Es kommen zwei Verteilung in Frage, nämlich die Normal- und die Gammaverteilung. Zunächst betrachten wir die Q-Q-Plots von `TEDY` verglichen mit den beiden Verteilungen, hier kommt das Paket `qualityTools` (siehe Roth, 2012) zur Anwendung, und ein Histogramm von der selbigen erklärenden Variable (siehe Abbildung 4.18).

```
> library(qualityTools)
> qqPlot(TEDY, "normal", confbounds = FALSE, cex.main=1.0, cex.lab=0.8)
> qqPlot(TEDY, "gamma", confbounds = FALSE, cex.main=1.0, cex.lab=0.8)
> hist(TEDY, cex.main=1.0, cex.lab=0.8)
```

Da die erzeugten Punkte im linken oberen Plot nicht näherungsweise auf einer Gerade liegen, kann eine Normalverteilung ausgeschlossen werden. Laut dem rechten Q-Q-Plot scheint jedoch die Annahme der Gammaverteilung plausibel zu sein. Auch das Histogramm verstärkt diese Mutmaßung. Sicherheitshalber werden noch Hypothesentests durchgeführt. Mithilfe des Shapiro-Wilk-Tests überprüfen wir zuerst die zugrunde liegende Stichprobe auf die Normalverteilung, anschließend unter Verwendung der `gammadist.test()`-Funktion des Pakets `goft` (siehe Gonzalez-Estrada und Villasenor-Alva, 2015) auf die Gammaverteilung.

```
> shapiro.test(TEDY)

Shapiro-Wilk normality test

data:  TEDY
W = 0.906, p-value < 2.2e-16

> library(goft)
```

```
> gammadist.test(TEDY)

Test of fit for the Gamma distribution

data:  TEDY
V = 0.8188, p-value = 0.5626
```

Der obige Output bestätigt unsere Vermutung. Der erste Test liefert einen signifikanten p-Wert, sodass die Annahme der Normalverteilung verworfen wird. Der zweite Test hingegen

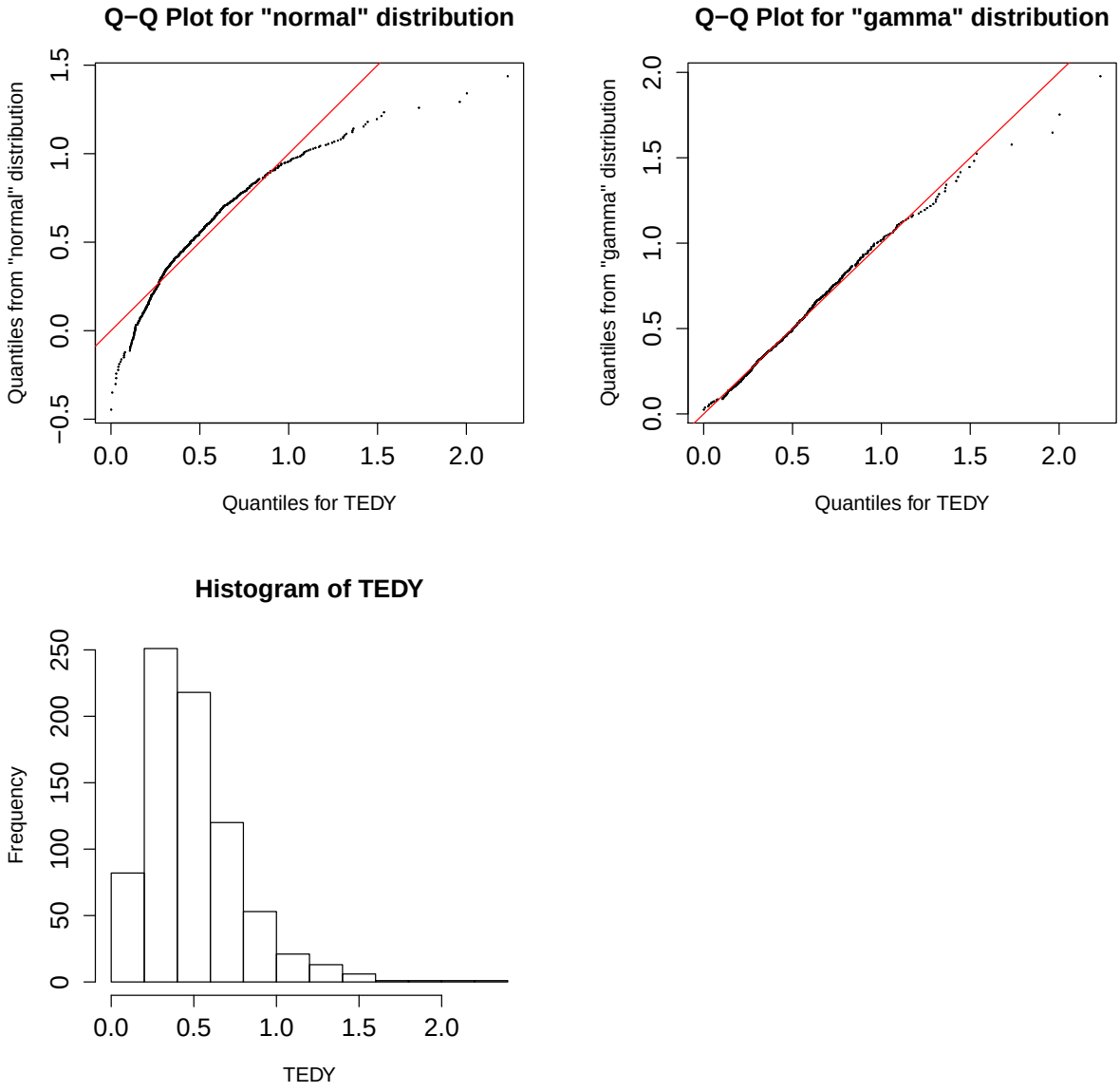

Abbildung 4.18: Oben die Q-Q-Plots von `TEDY` verglichen mit der Normal- bzw. Gammaverteilung und unten ein Histogramm von `TEDY`.

verwirft die Nullhypothese nicht, d.h. wir nehmen in der Folge an, dass die Daten gamma-verteilt sind.

Nach der Bestimmung der Verteilung ist es nun möglich, ein geeignetes Modell für `TEDY` zu erstellen. Wir versuchen zuerst eine Kubische Spline Basis mit `k` = 10 für den Glättungsterm der erklärenden Variable `time` und fügen dem Modell zusätzlich einen länderspezifischen Faktor hinzu. Für die Linkfunktion gebrauchen wir den Loglink.

```
> mod.time.TEDY <- gam(TEDY~s(time,bs="cr",k=10)+f_country,
+                      family=Gamma(linkv="log"))
> anova(mod.time.TEDY)

Family: Gamma
Link function: log

Formula:
TEDY ~ s(time, bs = "cr", k = 10) + f_country

Parametric Terms:
          df     F p-value
f_country 29 58.58  <2e-16

Approximate significance of smooth terms:
          edf Ref.df    F p-value
s(time) 6.640  7.745 38.8  <2e-16
```

Wie wir sehen können, liefert der `anova()`-Befehl sowohl für den parametrischen Term `f_country` als auch für `s(time)` einen äußerst signifikanten p-Wert, ebenso scheint die gewählte Basisdimension wegen den effektiven Freiheitsgraden von 6.640 in Ordnung zu sein. Wir haben also ein Modell gefunden, welches für jedes Land und jedes Jahr der Periode 1975 bis 2002 den Erwartungswert des Prädiktors schätzt. Gleichzeitig kann jetzt mithilfe einer sogenannten out-of-sample Prognose der Vorhersagewert für das Jahr 2003 kalkuliert werden. Um nun konkrete Werte zu erhalten, wird im Folgenden für zwei Länder, Jordanien und Türkei, eine Vorhersage erstellt, wobei der gewünschte Graph in Abbildung 4.19 dargestellt wird.

```
> # Prognose TEDY (Jordanien)
> pred_time <- seq(1, 29, length=1000)
> pred_factor_JO <- rep("JO", 1000)
> pred_factor_JO <- as.factor(pred_factor_JO)
> time_JO_data <- data.frame(time=pred_time, f_country=pred_factor_JO)
> pred_JO <- predict.gam(mod.time.TEDY, time_JO_data, se.fit=TRUE,
+                        type="response")
> plot(pred_time, pred_JO$fit, main="Prognose TEDY (Jordanien)", xlab="time",
+      ylab="Vorhersagewert", cex.main=1.0, cex.lab=0.8, cex=0.1)
```

Demzufolge unterscheiden sich die geschätzten Erwartungswerte der beiden Länder erheblich, jedoch kann ein identischer kubischer Verlauf beobachtet werden, wobei die Höchstwerte zwischen den Jahren 1987 und 1988 erreicht werden. Außerdem prognostiziert unser

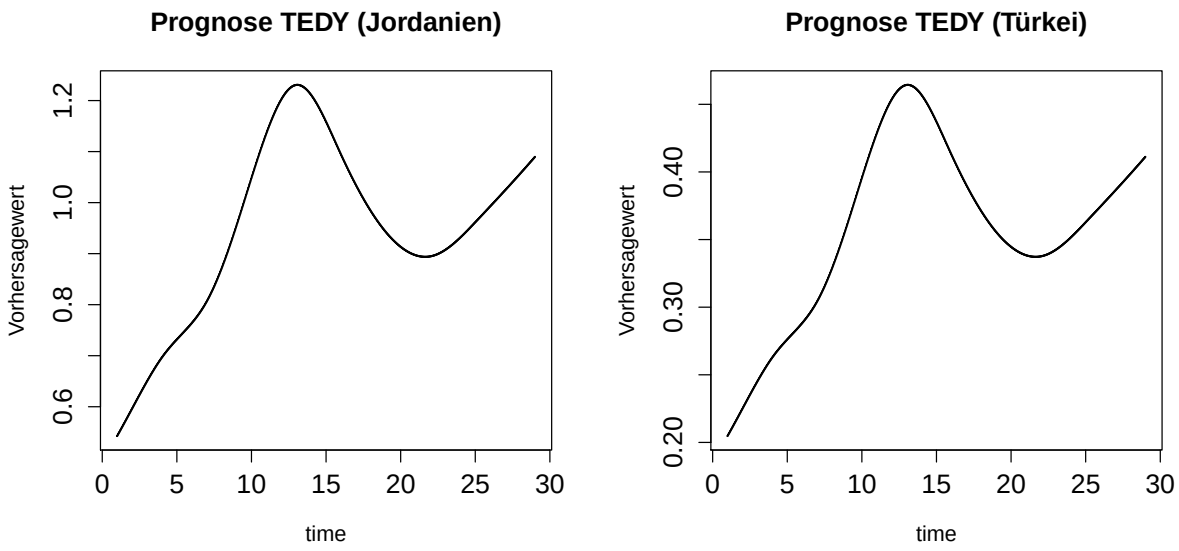

Abbildung 4.19: Links der geschätzte Erwartungswert des Prädiktors `TEDY` für Jordanien, rechts für die Türkei.

Modell einen Anstieg von `TEDY` für die Periode 2002 bis 2003, der exakte Wert der Vorhersage für 2003 (`time` = 29) beträgt 1.0896 bzw. 0.4111. Interessant wäre nun zu wissen, wie gut diese Prognose ist. Dafür betrachten wir den tatsächlichen Wert von `TEDY` im Jahr 2003. Dieser beläuft sich auf 1.2411 für Jordanien bzw. 0.4707 für die Türkei. Wie man sieht, sind die Abweichungen nur äußerst gering, somit scheint eine passende Vorhersage für `TEDY` mit unserem Modell möglich zu sein.

Für die restlichen zwei Prädiktoren, `EXCHRV` und `TORES`, wird nach dem obigen Schema ebenfalls ein Modell eruiert, wobei hier nur die generierten Graphen (siehe Abbildung 4.20 und 4.21) der geschätzten Erwartungswerte für Jordanien und die Türkei diskutiert werden. Genauere Details über das jeweilige Modell können im Anhang A.1 nachgelesen werden. Für `EXCHRV` lässt sich anhand Abbildung 4.20 wieder ein kubischer Verlauf ausmachen, mit Extremwerten ebenfalls in den Jahren 1987 und 1988. Ferner wird ein Abfall für die Jahre 2002 bis 2003 vorausgesagt, wobei sich der genaue Prognosewert für die beiden Länder im Jahr 2003 auf 0.0296 bzw. 0.3061 beläuft. Diese sind in der Größenordnung wieder korrekt, denn der tatsächliche Wert in 2003 beträgt 0.00001 bzw. 0.3411. Abschließend wird noch die Vorhersage von `TORES` erörtert. Im Unterschied zu den vorigen Prädiktoren ist ein völlig anderer Verlauf zu beobachten (siehe Abbildung 4.21). Für beide Länder besitzt der geschätzte Erwartungswert einen leicht wackeligen Anstieg, wobei ab dem Jahr 1989 (`time` = 15) eine deutliche Erhöhung dieser Steigung zu verbuchen ist und der Höchstwert mit 5 056 241 762 bzw. 22 548 499 146 im Jahr 2003 erreicht wird. Verglichen mit dem tatsächlichen Wert der WDI-Variable „Total reserves (includes gold)“, dieser beträgt 5 365 764 578 bzw. 35 548 509 248 im Jahre 2003, scheint unser Modell die Realität zufriedenstellend wiederzugeben, obwohl für die Türkei eine leichte Überschätzung vorliegt.

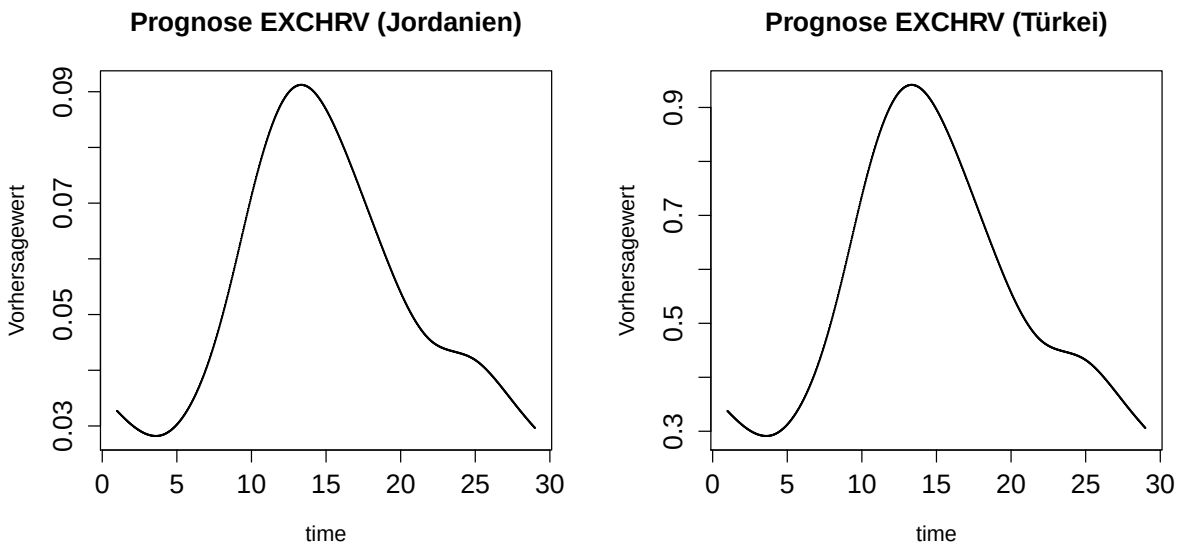

Abbildung 4.20: Links der geschätzte Erwartungswert des Prädiktors `EXCHRV` für Jordanien, rechts für die Türkei.

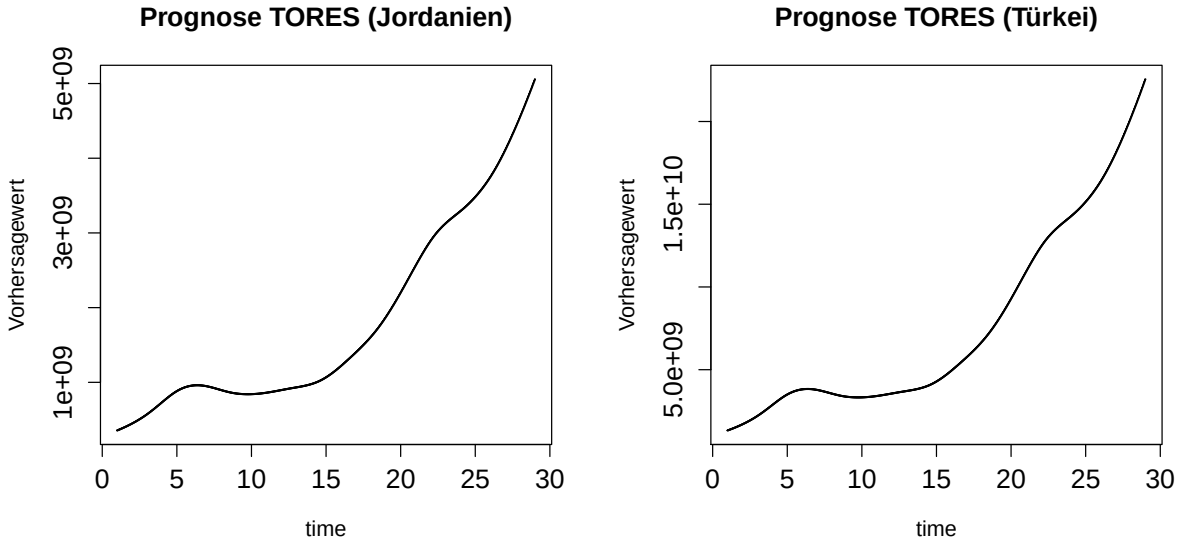

Abbildung 4.21: Links der geschätzte Erwartungswert des Prädiktors `TORES` für Jordanien, rechts für die Türkei.

Insgesamt kann festgehalten werden, dass für die Länder Jordanien und die Türkei mittels unseren erstellten Modellen passable Schätzungen der drei Prädiktoren möglich sind und somit gute Vorhersagen getätigt werden können. Nun erweitern wir unsere Prognosen und versuchen für die restlichen Länder den Erwartungswert der erklärenden Variablen für

Länder	TEDY	EXCHRV	TORES
BO	0.8738 (0.7168)	0.1620 (0.0934)	2 104 674 728 (1 096 941 362)
BR	0.3792 (0.4225)	0.4457 (0.2240)	62 356 397 636 (49 297 286 966)
CN	0.1323 (0.1252)	0.0537 (0.00009)	138 739 870 738 (416 199 410 849)
CO	0.3877 (0.3828)	0.1182 (0.1377)	19 426 608 706 (10 920 223 372)
CR	0.7014 (0.3224)	0.1042 (0.1129)	1 792 7866 965 (1 840 008 313)
DO	0.4205 (0.3420)	0.0579 (0.3293)	1 022 717 906 (279 543 950)
DZ	0.5930 (0.3420)	0.0848 (0.0234)	19 945 629 978 (35 454 600 240)
EG	0.8752 (0.3583)	0.0689 (0.2303)	15 823 198 218 (14 603 575 994)
GT	0.2663 (0.2353)	0.0439 (0.0095)	2 677 808 875 (2 924 966 747)
ID	0.6174 (0.5698)	0.1042 (0.0919)	33 088 887 646 (36 256 203 610)
IN	0.2302 (0.1923)	0.0528 (0.0324)	54 414 674 605 (103 737 207 867)
JM	1.0011 (0.5941)	0.1099 (0.1306)	703 070 170 (1 194 860 262)
JO	1.0896 (1.2411)	0.0296 (0.00001)	5 056 241 762 (5 365 764 578)
KZ	0.3549 (0.7530)	0.1679 (0.0318)	2 955 865 420 (4 962 118 090)
MA	0.7757 (0.3679)	0.0486 (0.0712)	5 551 650 492 (14 146 568 750)
MX	0.4529 (0.2220)	0.1566 (0.0678)	37 397 463 252 (59 026 700 138)
MY	0.4998 (0.4607)	0.0349 (0.0000)	41 894 451 453 (44 309 888 341)
PK	0.5673 (0.4380)	0.0487 (0.0604)	6 054 954 811 (11 815 604 851)
PA	0.9969 (0.5826)	$6 \cdot 10^{-6}$ (0.0000)	868 348 441 (1 010 969 628)
PE	0.7674 (0.5108)	0.3031 (0.0049)	10 892 212 872 (10 241 971 443)
PH	0.7465 (0.7480)	0.0574 (0.0850)	14 891 869 487 (17 083 609 581)
PY	0.4146 (0.6022)	0.0808 (0.2773)	2 121 880 109 (983 433 609)
RO	0.2103 (0.3798)	0.2242 (0.1841)	8 091 363 511 (9 449 445 049)
SV	0.4345 (0.5384)	0.0246 (0.0003)	2 011 304 984 (1 986 451 861)
TH	0.4809 (0.3839)	0.0300 (0.0441)	36 729 330 539 (42 161 780 744)
TN	0.6275 (0.6596)	0.0469 (0.0489)	3 208 050 357 (3 036 243 867)
TR	0.4111 (0.4707)	0.3061 (0.3411)	22 548 499 146 (35 548 509 248)
UA	0.3180 (0.5136)	0.3062 (0.0097)	1 940 237 600 (6 946 076 991)
ZA	0.2189 (0.2119)	0.0811 (0.1874)	14 626 200 444 (8 154 088 985)

Tabelle 4.8: Der geschätzte Erwartungswert der Prädiktoren für das Jahr 2003, in Klammer der tatsächliche Wert.

das Jahr 2003 zu bestimmen. Wichtig zu erwähnen ist hier obendrein, dass von den 47 Länder der Tabelle 4.1 nur noch für 29 Vorhersagewerte kalkuliert werden können, da die anderen Staaten aufgrund fehlender Beobachtungen in den von uns verwendeten Datenbanken nicht in die Modellschätzung eingebunden werden und folglich keine weitere Berücksichtigung finden. Die ermittelten Prognosen der 29 verbleibenden Länder werden in der Tabelle 4.8 dargestellt, wobei in der Klammer der tatsächliche Wert des Jahres 2003 angegeben wird. Wie auch schon für Jordanien und die Türkei liefern unsere Modelle im Großen und Ganzen gute Ergebnisse, obwohl die Vorhersagen natürlich nicht ganz exakt sind. Dabei sind besonders die Prognosen für `TEDY` und in Abstrichen für `EXCHRV` sehr erfreulich, aber auch für `TORES` sind viele der Schätzungen recht annehmbar. Nichts-

destotrotz sind für Letztere auch einige Unter- als auch Überschätzungen zu beobachten, die aber wahrscheinlich bei Werten dieser Größenordnung nicht allzu schwerwiegend sind. Zum jetzigen Zeitpunkt kann also noch nicht errechnet werden, ob diese Fehlschätzungen in Hinblick auf die Vorhersage der Krisen zu Problemen führen werden.

4.4.2 Vorhersage der Staatsschuldenkrisen

Nachdem die Vorhersagewerte der Prädiktoren für das Jahr 2003 ermittelt wurden, können wir nun diese in unser finales Modell einsetzen, um somit die prognostizierte Ausfallwahrscheinlichkeiten der Länder zu bestimmen. Die dieserart erstellten Prognosen vergleichen wir anschließend mit den tatsächlichen Werten, dafür wird wieder die CRAG-Datenbank der Bank of Canada herangezogen, und untersuchen, ob Vorhersagefehler vorliegen. Für uns ist es in der Folge sehr wichtig, dass so wenig Staatsschuldenkrisen wie möglich unerkannt bleiben, da ein Land in einer Schuldenkrise mit weit größeren Problemen zu kämpfen hat als ein Land, welcher von solcher nicht erfasst wird. Aus diesem Grund verwenden wir die folgende Spezifikation: Wird für die Insolvenz eines Landes eine Wahrscheinlichkeit von $\geq 42\%$ vorausgesagt, so nehmen wir an, dass eine Krise vorliegt. Nur für Wahrscheinlichkeiten unter 42% wird das Land als krisenfrei klassifiziert.

Die geschätzten Wahrscheinlichkeiten für das Auftreten einer Staatsschuldenkrise im Jahr 2003 für die 29 übrig gebliebenen Länder werden in der Tabelle 4.9 angegeben, ebenso wie die beobachteten Werte (0 = jeweiliges Land nicht insolvent, 1 = jeweiliges Land insolvent), und die Akkuratesse der Prognose. Dabei bezeichnen wir mit ✓ die Richtigkeit der Vorhersage und mit × die Unkorrektheit. Wie man sieht, erhalten wir zweifelsohne gute Ergebnisse, genauer gesagt stimmen 22 Mal Prognose und Realität überein, nur 7 Mal kommt es zu einem Vorhersagefehler. Auch werden lediglich 2 Krisen nicht aufgespürt, nämlich die in Indonesien und Paraguay. Ein Problem ist jedoch, dass bei einer Inkorrektheit der Prognose in einigen Fällen eine ziemlich große Abweichung vorliegt. Damit ist gemeint, dass obwohl sich laut der CRAG-Datenbank ein Land in einer Krise befindet, wir mit großer Wahrscheinlichkeit den komplementären Zustand prognostizieren, und umgekehrt. Nun wäre es aufschlussreich zu wissen, unter welchen Umständen diese Fehler auftreten und begeben uns deshalb auf Ursachenforschung.

Ursachen der Vorhersagefehler
Da die Gründe der Verfehlungen sehr unterschiedlich sind, betrachten wir jeden Fehler einzeln.

- Bolivien: Für Bolivien kann keine eindeutige Ursache der inkorrekten Prognose aufgespürt werden. Obwohl die Vorhersagen der Prädiktoren recht zufrieden stellend sind und die Reserven des Landes sogar überschätzt werden, prognostiziert unser Modell mit hoher Wahrscheinlichkeit eine Krise. Eine Vermutung ist, dass sich das Land noch nicht vollständig von den vorangegangenen Krisen, Bolivien ist von 1975 bis 2000 durchgehend insolvent, erholt hat und aus diesem Grund so wenig Reserven

Länder	Prognose	Realität	Akkuratesse
BO	0.7940	0	×
BR	0.6715	0	×
CN	$7.5 \cdot 10^{-12}$	0	✓
CO	0.0384	0	✓
CR	0.4946	1	✓
DO	0.7372	1	✓
DZ	0.1174	0	✓
EG	0.7405	1	✓
GT	0.1799	0	✓
ID	0.1605	1	×
IN	0.4055	0	✓
JM	0.9972	1	✓
JO	0.9982	1	✓
KZ	0.0950	0	✓
MA	0.4768	1	✓
MX	0.1165	0	✓
MY	0.1230	0	✓
PK	0.4027	0	✓
PA	0.9921	0	×
PE	0.9337	1	✓
PH	0.5768	0	×
PY	0.2053	1	×
RO	0.4295	1	✓
SV	0.1219	0	✓
TH	0.0536	0	✓
TN	0.2138	0	✓
TR	0.1026	0	✓
UA	0.7073	0	×
ZA	0.4051	0	✓

Tabelle 4.9: Die geschätzte Ausfallwahrscheinlichkeit des jeweiligen Landes für das Jahr 2003, der beobachtete Wert und die Korrektheit der Vorhersage, falls zur Klassifikation $\hat{\pi} \geq 0.42$ verwendet wird.

bzw. eine hohe Auslandsverschuldung aufweist. Unser Modell kann dies natürlich nicht berücksichtigen und sagt einen Kollaps voraus.

- Brasilien: Der Grund der fehlerhaften Vorhersage liegt hauptsächlich an der Überschätzung von `EXCHRV`. Wie schon Bolivien kämpft auch Brasilien bis 2000 mit enormen finanziellen Problemen, sodass sich der Wechselkurs erst im Anschluss stabilisieren kann. Infolgedessen wird für den Prädiktor ein zu hoher Wert geschätzt, welcher zu der inkorrekten Prognose der Eintrittswahrscheinlichkeit einer Insolvenz führt.

- Indonesien: Die laut der CRAG-Datenbank vorhandene Krise wird von unserem Modell nicht aufgespürt, es wird lediglich eine Ausfallwahrscheinlichkeit von etwa

16% vorausgesagt. Der Fehler kann nicht an der Prognose der Prädiktoren ausgemacht werden, denn deren Abweichung ist nur äußerst gering. Werden nun die Vorhersage- bzw. realen Werte der erklärenden Variablen näher betrachtet, so scheint die Schätzung unseres Modells angemessen zu sein, sodass diese Insolvenz als Sonderfall eingestuft werden kann.

- Panama: Über die Ursachen des Vorhersagefehlers kann nur spekuliert werden, denn auch mit den beobachteten Prädiktorwerten wird eine inkorrekte Voraussage erstellt. Ein möglicher Grund ist, wie bereits im Falle Boliviens und Brasiliens, dass sich das Land bis kurz vor 2003 in sehr großen finanziellen Schwierigkeiten befindet und somit noch nicht die nötigen Reserven aufbauen bzw. die Schulden verringern konnte.

- Philippinen: Auch hier ist die Frage nach den Ursachen der fehlerhaften Schätzung nur schwer zu beantworten, da die Prognose der erklärenden Variablen sehr gut ist. Die Vorhersage- bzw. tatsächlichen Werte deuten sowohl auf eine krisenbehaftete (siehe `TEDY`) als auch krisenfreie Zeit (siehe `EXCHRV` bzw. `TORES`) hin, sodass unser Modell die nachvollziehbare Ausfallwahrscheinlichkeit von 57.68% prognostiziert.

- Paraguay: Der Grund der falschen Prognose kann eindeutig an der Überschätzung von `TORES` festgemacht werden. Da Paraguay einen sehr atypischen Reserveverlauf aufweist, im Gegensatz zu den meisten anderen Ländern schrumpft der Wert von `TORES` am Ende der betrachteten Zeit, wird die Reserve des Landes überschätzt. Aufgrund des inkorrekten Prädiktorwertes wird im Anschluss auch die Wahrscheinlichkeit einer Krise falsch kalkuliert

- Ukraine: Die fehlerhafte Prognose der Ausfallwahrscheinlichkeit für dieses Land kann hauptsächlich an der Unterschätzung von `TORES` bzw. Überschätzung von `EXCHRV` ausgemacht werden. Wie schon Bolivien, Brasilien und Panama befindet sich auch die Ukraine bis kurz vor 2003 in einer Staatsschuldenkrise, genauer gesagt von 1993 bis 2002, sodass die Reserven verständlicherweise nicht allzu groß sind. Erst gegen Ende der krisenbehafteten Zeit steigen diese stark an und werden somit von unserem Modell unterschätzt. Desweiteren kann sich der Wechselkurs des Landes lediglich am Ende der finanziellen schwierigen Periode stabilisieren, sodass eine Überschätzung von `EXCHRV` die Folge ist. Aufgrund dieser beiden Fehlkalkulationen kommt es dann auch zu einer inkorrekten Schätzung der Eintrittswahrscheinlichkeit einer Krise.

Zusammenfassend kann gesagt werden, dass mittels unserem erstellten Modell die Eintrittswahrscheinlichkeit der Staatsschuldenkrisen von Schwellenländern gut modelliert werden kann, zudem gelingt es uns die meisten Krisen für das Jahr 2003 vorauszusagen. Dabei beträgt die Erfolgswahrscheinlichkeit unserer einjährigen Prognosen etwa 76%, ein ansehnlicher Wert. Die an dieser Stelle auftretenden Vorhersagefehler besitzen vielfältige und komplexe Ursachen, sodass es kompliziert ist, Verbesserungen für unser Modell abzuleiten. Eine Möglichkeit wäre jedoch sicherlich, „bessere“ Datensätze für unsere Schätzungen zu verwenden. Um die Reproduzierbarkeit dieser Arbeit sicherzustellen, haben wir bisher

sogenannte offene Daten verwendet, welche allerdings viele fehlende Beobachtungen aufweisen. Durch den Gebrauch von vollständigen, wahrscheinlich sehr kostspieligen Datenbanken könnte das Modell vermutlich um einiges verbessert werden. Weitere Alternativen wären z.B. auch die Aufnahme zusätzlicher Länder in die Stichprobe, die Ausdehnung der beobachteten Zeitperiode oder das Hinzufügen weiterer Indikatoren. Abschließend muss noch festgehalten werden, dass zuverlässige Ergebnisse bei Verwendung der obigen Modelle nur für kurzfristige Prognosen (z.B. einjährige) erwartet werden können, da eine Extrapolation in die fernere Zukunft viele Gefahren mit sich bringt (siehe Wood, 2006).

5 Resümee

Die Intention dieses Abschnitts ist die Zusammenfassung der Ergebnisse von Kapitel 4. Die Aufgabe bestand darin, mithilfe von Generalisierten Additiven Modellen die Eintrittswahrscheinlichkeit von Staatsschuldenkrisen in Schwellenländern zu modellieren sowie eine einjährige Vorhersage zu erstellen. Die Grundidee dafür lieferte die Arbeit von Manasse und Roubini (2005). In dieser wurden mithilfe der „Classification and Regression Tree“ (CART) Methodik 10 ökonomische bzw. politische Indikatoren festgestellt, welche einen signifikanten Einfluss auf die Zahlungsunfähigkeit eines Landes vorweisen, nämlich

- Total external debt over GDP (`TORES`),
- Short-term external debt over reserves ratio (`STDR`),
- Public external debt over revenue (`PEDR`),
- Real GDP growth (`RGRWT`),
- Inflation (`INF`),
- US treasury bill rate (`UST`),
- Exchange rate overvaluation (`EXCHRO`),
- Exchange rate variability (`EXCHRV`),
- Ratio to external financing requirements to foreign reserves (`FR`),
- Years to next presidential elections (`YNPRE`).

Unser Konzept war nun, diese 10 Merkmale als unsere Prädiktoren zu betrachten und so ein passendes Modell für die Ausfallwahrscheinlichkeit der Schwellenländer zu bestimmen. Doch bevor mit der Modellschätzung begonnen werden konnte, mussten passende Daten für die Responsevariable bzw. für die erklärenden Variablen gefunden werden. Hierbei entschieden wir uns jährliche Beobachtungen von 47 Ländern, diese stimmen mit jenen von Manasse und Roubini (2005) überein, der Periode 1975 bis 2002 für die Stichprobe zu gebrauchen. Darüber hinaus lag unser Hauptaugenmerk darauf, sogenannte offene Daten zu verwenden, sodass die Replizierbarkeit dieser Arbeit sichergestellt ist. Einige nennenswerte Quellen, welche schlussendlich zur Anwendung kamen, sind beispielsweise die CRAG-Datenbank der Bank of Canada und die World Developement Indicators (World Bank). Während erstere für die Responsevariable herangezogen wurde, kamen zweitere für die Prädiktoren zum Einsatz.

Nachdem die „bestmöglichen“ Daten ausgewählt wurden, richteten wir unsere Konzentration auf die Modellfindung. Die Vorgehensweise war dabei wie folgt: Zuerst betrachteten wir für jeden Prädiktor ein Modell mit einem Glättungsterm, kalkulierten die wichtigsten Kennzahlen und entschieden uns jenen signifikanten Prädiktor mit dem kleinsten p-Wert, natürlich unter Berücksichtigung der Sinnhaftigkeit dieser Entscheidung, in das Modell aufzunehmen. Nun überprüften wir, ob der Einfluss des ausgewählten Prädiktors parametrisch oder nichtparametrisch ist. Anschließend fügten wir einzeln die restlichen Prädiktoren dem Modell hinzu, testeten, welche zusätzlich signifikant sind und wählten anhand des kleinsten p-Werts den nächsten „optimalen“ Prädiktor aus. Diese Strategie wurde solange fortgeführt, bis keine signifikanten Prädiktoren mehr eingebunden werden konnten. Letztendlich erhielten wir ein Modell mit zwei erklärenden Variablen, nämlich `EXCHRV` und `TEDY`.

Zu unserer Überraschung fand jedoch der Prädiktor `STDR`, das Verhältnis der kurzfristigen Auslandsverschuldung zu den Reserven eines Landes, keine Berücksichtigung, sodass also die Reserven eines Landes in keinster Weise in unser Modell einflossen. Daher prüften wir, ob womöglich die WDI-Variable „Total reserves (includes gold)“ (`TORES`) relevant ist und eine Verbesserung darstellt. Das Ergebnis war sehr erfreulich, denn wir erhielten einen signifikanten p-Wert und ein deutlich verbessertes Modell. Abschließend wurde noch untersucht, ob das Hinzufügen eines länderspezifischen Faktors sinnvoll wäre. Es konnte festgestellt werden, dass mit einem Modell, bei welchem die Eintrittswahrscheinlichkeit einer Staatsschuldenkrise auch vom betrachteten Land abhängt, keine vernünftige Schätzung möglich ist. Deshalb wählten wir als finales Modell für den linearen Prädiktor

$$\hat{\eta} = \hat{f}_1(\texttt{TEDY}) + \hat{f}_2(\texttt{EXCHRV}) + \hat{f}_3(\texttt{TORES}) + 0.7763.$$

Die geschätzten Funktionen der drei Prädiktoren wurden in der Abbildung 4.17 dargestellt. Wichtig ist es hier noch anzuführen, dass Beobachtungen von 29 Ländern in die Schätzung des finalen Modells eingegangen sind. Aufgrund unvollständiger Datensätze fanden die übrigen Staaten keine Berücksichtigung.

Im nächsten Schritt versuchten wir die Ausfallwahrscheinlichkeiten der Schwellenländer für das Jahr 2003 zu prognostizieren. Dabei stellte sich das Problem, dass unser Modell auf die Werte der drei Prädiktoren angewiesen ist, welche jedoch im Jahr 2002 noch nicht bekannt sind. Aus diesem Grund bestimmten wir mithilfe von GAMs ein Modell, welches bezüglich der Zeit den Erwartungswert der jeweiligen erklärenden Variable schätzt. Überdies wurde ein länderspezifischer Faktor hinzugefügt. Nun war es möglich für jedes Land eine Vorhersage für den jeweiligen Prädiktor für das Jahr 2003 zu generieren, wobei die Resultate im Großen und Ganzen zufriedenstellend waren. Verglichen mit den tatsächlichen Werten waren insbesondere die Prognosen für `TEDY` und in Abzügen für `EXCHRV` ansehnlich. Darüber hinaus waren auch viele der Schätzungen von `TORES` passabel, obwohl hier vereinzelt größere Unter- bzw. Überschätzungen vorlagen, welchen aber aufgrund der Größenordnung der erklärenden Variable keine allzu große Beachtung geschenkt wurde. Nachdem die Vorhersagewerte der Prädiktoren für die jeweiligen Länder eruiert wurden, fügten wir diese in das finale Modell, `mod.final`, ein, um so die Eintrittswahrscheinlichkeit der Staatsschuldenkrisen für 2003 zu ermitteln. Die resultierenden Ergebnisse waren sehr beachtlich, denn etwa 76% der einjährigen Prognose waren korrekt. Darüber hinaus konnten bis auf zwei Ausnahmen alle Krisen vorhergesagt werden. Zum Abschluss wurden noch die Ursachen der Vorhersagefehler analysiert, welche jedoch mannigfaltig waren, sodass ein wiederholter Grund nur schwer ausgemacht werden konnte.

Insgesamt konnte also festgestellt werden, dass insbesondere die Indikatoren `TEDY`, `EXCHRV` und `TORES` für die finanzielle Stabilität eines Landes verantwortlich sind. Unsere Prognosen sind beachtlich, jedoch würde sich noch eine Ausweitung unserer Analysen anbieten. Durch das Inkludieren zusätzlicher Länder in die Stichprobe, die Vergrößerung der betrachteten Jahre oder die Hinzunahme anderer Indikatoren könnte unser Modell wahrscheinlich weiter verbessert werden. Ein interessanter Versuch wäre es zudem, zufällige Effekte in das Modell hinzuzufügen.

A Anhang

A.1 Vorhersage von EXCHRV und TORES

Die Bestimmung der Vorhersage für die zwei Prädiktoren erfolgt äquivalent nach dem Schema, welches bereits für TEDY in Abschnitt 4.4.1 verwendet wurde. Jedoch kommen einige Schwierigkeiten auf, sodass eine nähere Betrachtung sinnvoll ist. Wir diskutieren ausschließlich die mithilfe von GAMs erstellten Modelle der zwei erklärenden Variablen, die darauffolgenden out-of-sample Prognosen geschehen analog zu Abschnitt 4.4.1, wobei die Ergebnisse für das Jahr 2003 in der Tabelle 4.8 zu finden sind.

Modell für EXCHRV
Im ersten Schritt muss die Verteilung des Prädiktors festgelegt werden. Wir untersuchen zuerst auf eine Normalverteilung und betrachten dafür den entsprechenden Q-Q-Plot von EXCHRV und ein Histogramm (siehe Abbildung A.1). Es ist offensichtlich, dass die Annah-

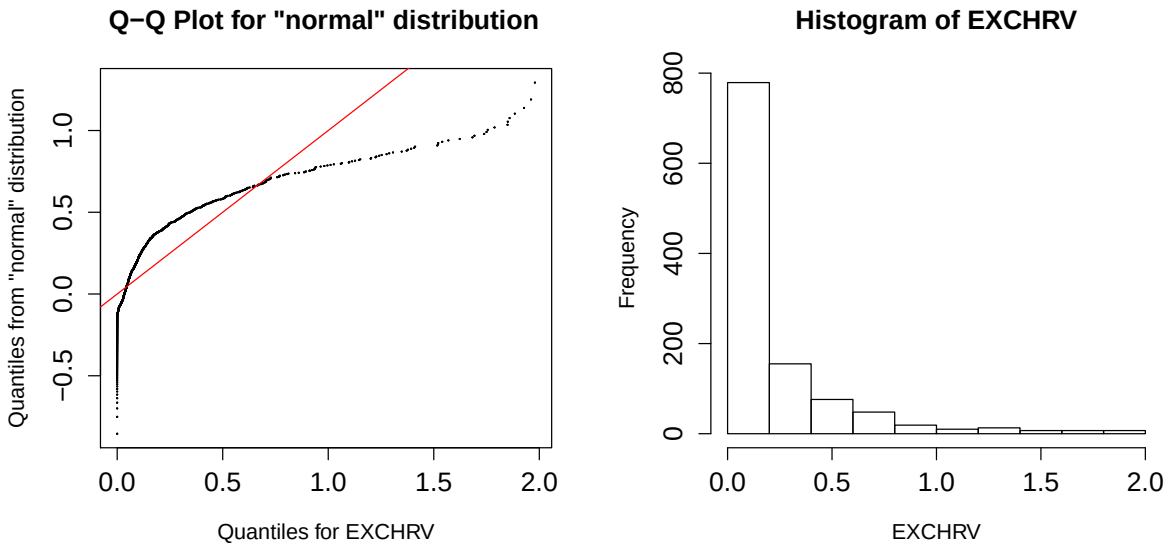

Abbildung A.1: Links der Q-Q-Plot von EXCHRV verglichen mit der Normalverteilung und rechts ein Histogramm von EXCHRV.

me der Normalverteilung verworfen werden kann. Aufgrund des Histogramms scheint die Gammaverteilung ein möglicher Kandidat zu sein. Das Problem, welches nun auftritt, ist, dass EXCHRV eine nicht unerhebliche Menge von Nullwerten innehat, jedoch die Gammaverteilung nur für Werte größer als Null definiert ist. Aus diesem Grund verschieben wir alle EXCHRV-Werte um 0.00001, sodass die Nullwerte der Stichprobe eliminiert werden, erstellen mit diesen das Modell bzw. die Vorhersage und transformieren abschließend wieder zurück. Durch diese Vorgangsweise ändert sich die Prognose nur marginal. Würde man jedoch alle Nullwerte entfernen und die vorhandene Information nicht verwerten, so ist das Risiko einer falschen Vorhersage erheblich größer. Für die neuen EXCHRV-Werte, EXCHRV.new, testen wir

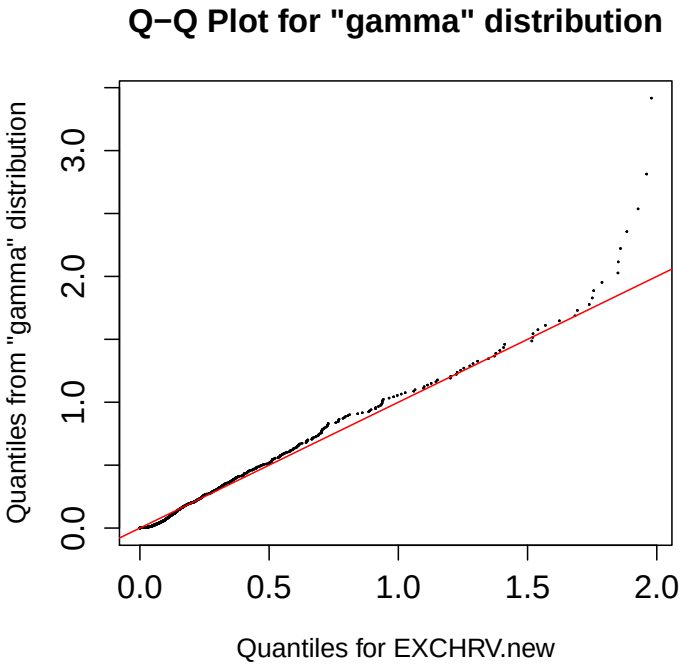

Abbildung A.2: Der Q-Q-Plot von `EXCHRV.new` verglichen mit der Gammaverteilung.

also, ob die Gammaverteilung tatsächlich eine geeignete Wahl ist (siehe Abbildung A.2). Die Annahme scheint korrekt zu sein, denn nur für einige wenige Werte sind Abweichungen zu beobachten. Nachdem die Verteilung festgesetzt wurde, können wir ein Modell für den Prädiktor bestimmen. Unser erster Ansatz ist der Folgende: Wir gebrauchen den Loglink, verwenden eine Kubische Spline Basis mit `k` = 10 für den Glättungsterm des Prädiktors `time` und ergänzen das Modell mit einem länderspezifischen Faktor:

```
> mod.time.EXCHRV <- gam(EXCHRV.new~s(time,bs="cr",k=10)+f_country,
+                        family=Gamma(link="log"))
> anova(mod.time.EXCHRV)

Family: Gamma
Link function: log
Formula:
EXCHRV.new ~ s(time, bs = "cr", k = 10) + f_country

Parametric Terms:
          df     F p-value
f_country 45 71.34  <2e-16

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(time) 5.649  6.775 23.14  <2e-16
```

Bei Betrachtung des obigen Outputs scheint das generierte Modell in Ordnung zu sein, denn neben dem Glättungsterm `s(time)` weist auch der parametrische Term `f_country` einen höchst signifikanten p-Wert auf; ebenso ist die gewählte Basisdimension von 10 angemessen. Also ist unser Modell für `EXCHRV` bestimmt. Unter Berücksichtigung der Rücktransformation können nun die Prognosen für das Jahr 2003 sehr leicht ermittelt werden.

Modell für `TORES`
Auch hier sind wieder einige Schwierigkeiten zu beobachten, sodass eine genauere Analyse unumgänglich ist. Ein Problem ist, dass die Werte von `TORES` extrem groß sind, wobei Reserven über 10 Milliarden keine Seltenheit sind. Vor diesem Hintergrund logarithmieren wir die Werte des Prädiktors. Nun betrachten wir die entsprechenden Q-Q-Plots für `log(TORES)` und ein Histogramm, um die Verteilung des Prädiktors zu bestimmen (siehe Abbildung A.3). Es scheinen beide Verteilung geeignet zu sein, aufgrund der leichten Schiefe im Histogramm kommt jedoch eher eine Gammaverteilung in Frage. Um sicher zu gehen, führen wir noch Hypothesentests durch, wobei der Shapiro-Wilk-Test für die Normalverteilung und die `gammadist.test()`-Funktion des Paktes `goft` (siehe Gonzalez-Estrada und Villasenor-Alva, 2015) für die Gammaverteilung zur Anwendung kommt.

```
> shapiro.test(log(TORES))

Shapiro-Wilk normality test
data:  log(TORES)
W = 0.9956, p-value = 0.002057

> gammadist.test(na.omit(log(TORES)))

Test of fit for the Gamma distribution
data:  na.omit(log(TORES))
V = -0.6796, p-value = 0.6309
```

Wie man sieht, wird unsere Vermutung durch die beiden Tests bestätigt. Also wird für die weiteren Analysen die Gammaverteilung angenommen. Nachdem die Verteilung geklärt wurde, versuchen wir einige Ansätze um ein passendes Modell für den Prädiktor zu finden. Letztendlich entscheiden wir uns für das folgende Modell.

```
> mod.time.TORES <- gam(log(TORES)~s(time,bs="cr",k=15)+f_country,
+                       family=Gamma(link="log"))
> anova(mod.time.TORES)

Family: Gamma
Link function: log
Formula:
log(TORES) ~ s(time, bs = "cr", k = 15) + f_country
Parametric Terms:
          df     F p-value
f_country 46 163.4  <2e-16
```

```
Approximate significance of smooth terms:
           edf Ref.df     F p-value
s(time)  9.149 10.898 203.1  <2e-16
```

Nun haben wir ein Modell für `log(TORES)` gefunden und können somit die logarithmierten Prognosen für das Jahr 2003 erstellen. Um die korrekten Vorhersagewerte zu erhalten, müssen die logarithmierten Werte natürlich wieder zurücktransformiert werden.

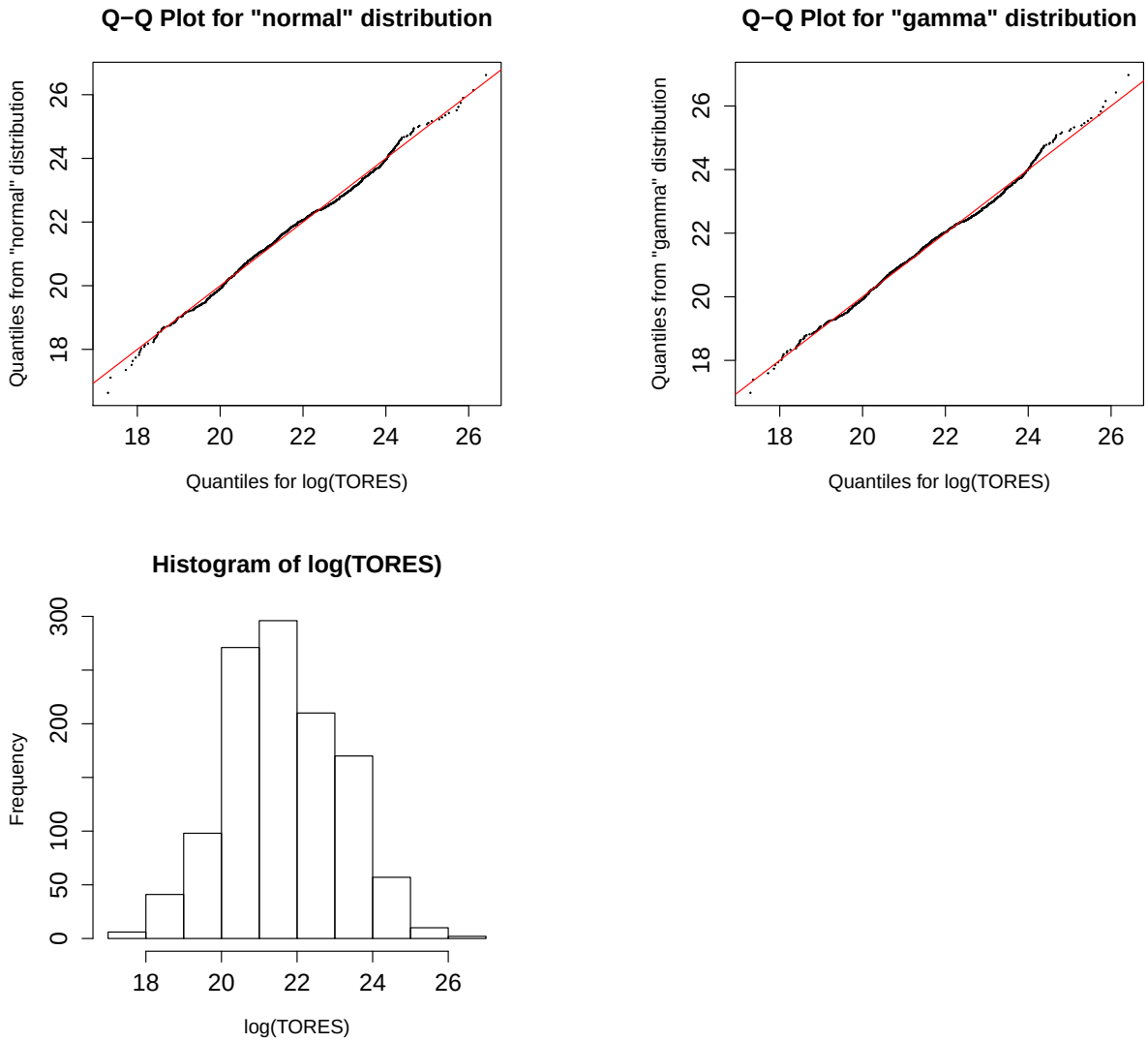

Abbildung A.3: Oben die Q-Q-Plots von `log(TORES)` verglichen mit der Normal- bzw. Gammaverteilung und unten ein Histogramm von `log(TORES)`.

Literatur

Beers, D. T. und Nadeau, J.-S. (2014). Introducing a new database of sovereign defaults. *Technical Report No. 101, Bank of Canada*.

Binder, H. und Tutz, G. (2008). A comparison of methods for the fitting of generalized additive models. *Statistics and Computing*, *18*, 87-99.

Buja, A., Hastie, T. und Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, *17* (2), 453-510.

Cox, M. (1972). The numerical evaluation of B-splines. *Journal of the Institute for Mathematics Applications*, *10*, 134-149.

Craven, P. und Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*, 377-403.

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models* (Zweite Aufl.). Boca Raton: Chapman and Hall/CRC.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In W. Schempp und K. Zeller (Hrsg.), *Constructive theory of functions of several variables* (S. 85-100). Berlin/Heidelberg: Springer.

Easterly, W. R. und Levine, R. (2002). Tropics, germs, and crops: how endowments influence economic development. *NBER Working Paper*, *9106*.

Eilers, P. H. C. und Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11* (2), 89-102.

Fahrmeir, L. und Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, *13* (1), 342-368.

Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*. Berlin/Heidelberg: Springer.

Golub, G. H., Heath, M. und Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21* (2).

Gonzalez-Estrada, E. und Villasenor-Alva, J. A. (2015). goft: Tests of fit for some probability distributions [Software-Handbuch]. Zugriff auf `http://CRAN.R-project.org/package=goft` (R package version 1.1)

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Gu, C. und Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing*, *12* (2), 383-398.

Hardin, J. W. und Hilbe, J. M. (2012). *Generalized Linear Models and Extensions* (Dritte Aufl.). College Station: Stata Press.

Hastie, T. und Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1* (3), 297-318.

Hastie, T. und Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall/CRC.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, *15* (3), 958-975.

Manasse, P. und Roubini, N. (2005). "Rules of thumb" for sovereign debt crises. *IMF Working Paper*, *05/42*.

Manasse, P. und Roubini, N. (2009). "Rules of thumb" for sovereign debt crises. *Journal of International Economics*, *78*, 192-205.

Marx, B. D. und Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, *28*, 193-209.

McCullagh, P. und Nelder, J. A. (1989). *Generalized Linear Models* (Zweite Aufl.). Boca Raton: Chapman and Hall/CRC.

Milicer, H. und Szczotka, F. (1966). Age at menarche in Warsaw girls in 1965. *Human Biology*, *38*, 199-203.

Nelder, J. A. und Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A (General)*, *135* (3), 370-384.

Patrikalakis, N. M. und Maekawa, T. (2002). *Shape Interrogation for Computer Aided Design and Manufacturing*. Berlin/Heidelberg: Springer.

Pokropp, F. (1994). *Lineare Regression und Varianzanalyse*. Berlin/Heidelberg: Oldenbourg Wissenschaftsverlag.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, *9* (4), 705-724.

R Core Team. (2013). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Zugriff auf `http://www.R-project.org/`

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik*, *10*, 177-183.

Roth, T. (2012). qualitytools: Statistics in quality science. [Software-Handbuch]. Zugriff auf `http://www.r-qualitytools.org` (R package version 1.54)

Ruppert, D., Wand, M. P. und Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schoenberg, I. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics*, *4*, 45-99.

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society, B*, *47* (1), 1-52.

Tutz, G. und Binder, H. (2006). Generalized additive models with implicit variable selection by likelihood-based boosting. *Biometrics*, *62* (4), 961-971.

Venables, W. und Ripley, B. (2002). *Modern Applied Statistics with S-Plus* (Vierte Aufl.). New York: Springer.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society, C*, *31* (2), 144-148.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman and Hall/CRC.

Wood, S. N. (2016). Package "mgcv" [Software-Handbuch]. Zugriff auf `https://cran.r-project.org/web/packages/mgcv/mgcv.pdf`